Dipl.-Ing. Stefan Grabuschnig, BA

Discovery of disease related patterns in cell-free DNA

A framework for differential analysis and detection of potential biomarkers in cell-free DNA sequencing data

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der technischen Wissenschaften

eingereicht an der

Technischen Universität Graz

Betreuerin/Betreuer

Christoph W. Sensen, Univ.-Prof. Dipl.-Biol. Dr.rer.nat.
Institute for Computational Biotechnology
Doctoral School of Molecular Biosciences and Biotechnology

Graz, June 2021

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

____

Datum, Unterschrift

GRAZ UNIVERSITY OF TECHNOLOGY

Abstract

Faculty of Technical Chemistry, Chemical and Process Engineering, Biotechnology
Institute of Computational Biotechnology

Doctor of Technical Sciences

Discovery of disease related patterns in cell-free DNA

by DI Stefan GRABUSCHNIG

Cell-free DNAs (cfDNAs) are mainly short fragments of DNA which are found in all vertebrate body fluids. Their concentration in the blood stream increases in connection to various disease conditions. This work introduces a dedicated approach to the efficient analysis of cfDNA sequencing data with the aim of detecting sequence motifs with altered levels of occurrence in relation to disease conditions. We have successfully used such motifs as biomarkers for the creation of an diagnostic assay for human sepsis, which has the potential to outperform the current diagnostic standards. We also investigated the composition of blood cell-free DNA of healthy and diseased donors, where we found that certain species of satellite DNA and retrotransposable elements (RTEs) are significantly overrepresented in the cfDNA population. Additionally, the fraction of RTEs was substantially increased in connection to disease conditions, which we observed for human as well as bovine donors. Since several biological mechanisms seem to be involved in the generation of cfDNA molecules, we aligned our results with the scientific literature in a review article about the potential origins of the cfDNA in human blood.

Acknowledgements

First of all I want to thank our Lab team Elisabeth Hirschböck, Thorsten Bachler, Dr. Petra Heidinger and also Daniel Schwendenwein for all their effort, good work and cooperation. I also want to thank Dr. Klaus Schliep, Dr. Petra Heidinger, Daniel Schwendenwein and Max Malek for proofreading this thesis. Special thanks go to Dr. Klaus Schliep for many inspiring conversations and encouragements, to Dr. Jung Soh and Max Malek for their excellent technical advices and to Daniela Starcevic for keeping our institute alive and running. Finally, I want to thank my mentor Prof. Dr. Christoph W. Sensen for giving me the opportunity to do my PhD under his supervision, who also gave me a lot of freedom in this matter and encouraged me to pursue my own path and ideas in our research.

Contents

Abstract iii

Acknowledgements v

1 Introduction 1

2 Analysis of cell-free DNA sequencing data 3
- 2.1 Objectives . . . 3
- 2.2 Properties of the data . . . 4
  - 2.2.1 Paired-end DNA sequencing data . . . 4
  - 2.2.2 Genomic Alignments . . . 5
  - 2.2.3 Count- and coverage data . . . 6
    - Fragment reconstruction . . . 7
    - Normalization . . . 7
- 2.3 Preprocessing: Efficient access to the data . . . 7
  - 2.3.1 Calculation of coverage data . . . 8
  - 2.3.2 Storing and compression . . . 8
- 2.4 Analysis of cfDNA coverage . . . 9
- 2.5 Identification of cfDNA biomarkers . . . 9
  - 2.5.1 Sliding window scan and de-noising . . . 10
  - 2.5.2 Outlier removal . . . 11
  - 2.5.3 Determination of candidate marker regions . . . 12
  - 2.5.4 Candidate marker evaluation . . . 12
- 2.6 Analysis of cfDNA composition and properties . . . 12

3 Overview of the biology and origins of cell-free DNA in humans 19
- 3.1 Prologue . . . 19
- 3.2 Article . . . 19
- 3.3 Author rights and permissions . . . 20

4 The composition of cell-free DNA in blood serum and plasma 45
- 4.1 Prologue . . . 45
- 4.2 Article . . . 45
- 4.3 Author rights and permissions . . . 46

5 Identification and application of disease related cfDNA biomarkers 61
- 5.1 Prologue . . . 61
- 5.2 Article . . . 62
- 5.3 Author rights and permissions . . . 62

6 Conclusion 79

Bibliography 83

List of Figures

2.1 Fragment reconstruction 14
2.2 Data preprocessing 15
2.3 Coverage plot 16
2.4 Marker prediction 17

List of Abbreviations

AA	**A**mino **A**cid
BAM	**B**inary **A**lignment **M**ap
BFB	**B**reakage **F**usion **B**ridge
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
bp	**b**ase-**p**air(s)
Kbp	**K**ilo-**b**ase-**p**airs
Mbp	**M**ega-**b**ase-**p**airs
Gbp	**G**iga-**b**ase-**p**airs
MB	**M**ega-**B**yte(s)
GB	**G**iga-**B**yte(s)
CAD	**C**aspase **A**ctivated **D**Nase
caret	**c**lassification **a**nd **re**gression **t**raining
centr	**cent**romeric alpha satellite **r**epeat
CIN	**C**hromosome **IN**stability
CNAs	**C**irculating **N**ucleic **A**cids
Cq	**q**uantitation **C**ycle
DFFB	**D**NA **F**ragmentation **F**actor subunit **B**eta
DMs	**D**ouble **M**inutes
DNA	**D**eoxyribo**N**ucleic **A**cid
cDNA	**c**omplementary **DNA**
cfDNA	**c**ell-**f**ree **DNA**
cffDNA	**c**ell-**f**ree **f**oetal **DNA**
ctDNA	**c**irculating-**t**umor **DNA**
dsDNA	**d**ouble-**s**tranded **DNA**
mtDNA	**mit**ochondrial **DNA**
DNase	**D**eoxyribo**N**uclease
DRDC	**D**efense **R**esearch and **D**evelopment **C**anada
DSBs	**D**ouble-**S**tranded DNA **B**reaks
EBI	**E**uropean **B**ioinformatics **Institute**
ENA	**E**uropean **N**ucleotide **A**rchive
EVs	**E**xtracellular **V**esicles
FDA	U.S. **F**ood and **D**rug **A**dministration
FSAP	plasma **F**actor **VII** **a**ctivating **P**rotease
FPKM	plasma **F**ragments **P**er **K**ilobase per **Million** mapped fragments
IC	**I**nterstitial **C**ystitis
k-MT	**k**inetochore **M**icro**T**ubules
LINE	**L**ong **I**nterspersed **N**uclear **E**lement
MANOVA	**M**ultivariate **AN**alysis **O**f **V**ariance
MAP	*Mycobacterium Avium ssp. Paratuberculosis*
MeDIP	**Me**thylated **D**NA **I**mmuno-**P**recipitation
MN	**M**icro **N**ucleus
MNi	**M**icro **N**ucle**i**
MSAs	**M**ultiple **S**equence **A**lignment

MVBs	**M**ulti**V**esicular **B**odies
MUG	**M**edical **U**niversity of **G**raz
NCBIs	**N**ational **C**enter for **B**iotechnology **I**nformation
NBUDs	**N**uclear **BUD**s
NETs	**N**eutrophile **E**xtracellular **T**raps
OMVs	**O**uter **M**embrane **V**esicles
OR	**O**ut of **R**each
PCR	**P**olymerase **C**hain **R**eaction
qPCR	**q**uantitative real-time **PCR**
PDB	**P**rotein **D**ata **B**ank
PRR	**P**attern **R**ecognition **R**eceptor
RBCs	**R**ed **B**lood **C**ells
RCSB	**R**esearch **C**ollaboratory for **S**tructural **B**ioinformatics
RNA	**R**ibo**N**ucleic **A**cid
mRNA	**m**essenger **RNA**
miRNA	**mi**cro**RNA**
RNA-Seq DE	**RNA Seq**uencing **D**ifferential **E**xpression
ROS	**R**eactive **O**xygen **S**pecies
RPKM	**R**eads **P**er **K**ilobase per **Million** mapped reads
RTE	**R**etro-**T**ransposable **E**lement
SAM	**S**equence **A**lignment **M**ap
SINE	**S**hort **I**nterspersed **N**uclear **E**lement
SVA	**S**INE/**V**NTR/**A**lu
UCSC	**U**niversity of **C**alifornia **S**anta **C**ruz
VNTR	**V**ariable **N**umber **T**andem **R**epeat

List of Symbols

i	sample index	1
c	chromosome number	1
p	nucleotide position	1
p^W	sliding window position	1
l	sliding window size	bp
S	sample set	nominal
$Rfam$	repeat family	nominal
μ	arithmetic mean	-
σ^2	variance	-
ϵ	offset	-
$cov_{Ai}(c, p)$	absolute coverage	1
$cov_{Ri}(c, p)$	relative (normalized) coverage	1
$cov_{Ri}^W(c, p^W)$	windowed relative (normalized) coverage	1
$z_i(c, p^W)$	z-score	1
$cov_{Ri}(RFam)$	relative coverage of repeat family	1
$abundance(Rfam)$	genomic abundance of repeat family	%
$rep(RFam)$	representation of repeat family	%
$depth_{Ri}(Rfam))$	relative coverage depth of repeat family	%
$COV_S(c, p)$	coverage distribution	1
$COV_S^W(c, p^W)$	windowed coverage distribution	1
KL	Kullback-Leibler Divergence	1

Dedicated to my parents and to my sister who always supported and encouraged me.

Chapter 1

Introduction

The history of computational Bioinformatics sequence analysis began after Pehr V. Edman's degradation method [1, 2] made the determination of amino acid (AA) sequences feasible [3, 4]. Since the Edman reaction only generated sequences with a length of about 50 to 60 residues in a single run [2], proteins usually had to be cleaved into smaller bits before sequencing. The final sequence of a protein had therefore to be assembled from the peptide subsequences obtained from many individual runs [4]. The assembly process was initially performed manually without the aid of computers, which was very tedious and took months [3]. This changed in the early sixties when Margaret Dayhoff developed the first Bioinformatics computer program COMPROTEIN [5], a *de novo* sequence assembly approach implemented in the Fortran programming language. The algorithm was entirely coded on punch cards and required mere minutes for solving the assembly task when executed on an IBM 7090 machine [4]. The subsequently increasing availability of AA sequences provided the basis for the realization of Emile Zuckerkandl's and Linus Pauling's idea to derive evidence for evolutionary history from interspecies differences between the AA sequences of proteins [3, 6, 7]. This feat massively extended the methodological possibilities of phylogenetic analyses. Formerly mainly phenotypical, biochemical and behavioral traits were used for the determination of phylogenetic relationships. Now these relations could be derived directly from observing the evolution of AA and DNA molecules [3, 4]. The progressive developments raised the demand for efficient means to compare AA sequences in order to assess similarity and homology, which has become one of the most commonly performed tasks in Bioinformatics [8]. Answering the need, Saul B. Needleman and Christopher D. Wunsch developed their famous algorithm for pairwise AA sequence alignment in 1970 [9] based on pioneer work from Walter M. Fitch [10]. Still, calculating a multiple sequence alignment (MSA) comprised a serious challenge. Since the number of iteratively constructed alignments depends exponentially on the number of sequences used [11], evaluating all possible alignments in order to find an optimal MSA was computationally impractical [4]. The issue was finally overcome by the approach of Da-Fei Feng and Russel F. Doolittle [12]. Their algorithm used a guide tree, precomputed from pairwise alignments, for the construction of a MSA [4] and provided the basis for Desmond D. Higgins' and Paul M. Sharp's popular CLUSTAL MSA algorithm [13] as well as many further developments [14].

The deciphering of the genetic code [15], initiated by J. Heinrich Matthaei's and Marshall W. Nirenberg's famed experiment [16], enabled the translation between DNA, mRNA and AA sequences corresponding to Francis H. C. Crick's central dogma of molecular biology [17]. The field of application for Bioinformatics sequence analysis then quickly expanded after the introduction of Fred Sanger's and Alan Coulson's chain termination method [18]. It was developed in succession to their own "plus and minus" [19] and Allen Maxam's and Walter Gilbert's method [20], and made sequencing of longer DNA molecules feasible. Up to now many different fields of application for the analysis of AA, DNA or RNA sequences have evolved, each coming with its own challenges and accordingly tailored methods [21]. While phylogenetic analyses [22], genome assemblies [23] or automated annotation tasks [24] predominantly constitute sequence comparison tasks, other major applications involve pattern recognition tasks, *e.g.* for motif discovery [25], or mapping and coverage quantification of RNA sequencing data [26]. The establishment of large biomolecular databases [27], such as the NCBI GenBank [28], EBI ENA [29] or the RCSB Protein Data Bank PDB [30], which can be efficiently queried with tools like BLAST [31], led to the nowadays modern Bioinformatics [4].

Bioinformatics analysis of circulating cell-free DNA (cfDNA) sequencing data is an emerging application of sequence analysis, coming with long-familiar problems but also with new challenges (as outlined in Chapter 2) [32, 33, 34]. This is partly related to the fact that many aspects of the biology behind the occurrence of cfDNA molecules are currently not yet fully understood. These extracellular DNA fragments are present in vertebrate body-fluids in various forms [35, 36]. They were initially discovered by Mandel and Metais in 1948 [37] even before the structure of the DNA double-helix was solved by Watson and Crick [38]. cfDNA molecules originate from a number of active and passive release mechanisms, which we reviewed in our article in Chapter 3 of this work. Elevated cfDNA concentrations in the bloodstream were reported for several disease conditions [37] such as cancer [39], rheumatoid arthritis [40] or autoimmune diseases [41]. These observations led to a keen interest in using cfDNA as non-invasive diagnostic biomarker in the form of so called "liquid biopsies" [42, 43]. In order to avoid redundancies, I abdicate from going into much detail within this introduction since already two chapters of this work are dedicated to the biological mechanisms behind the occurrence of cfDNAs in the bloodstream (see Chapter 3) and their properties (see Chapter 4).

This thesis is dedicated to the Bioinformatics analysis of cfDNA sequencing data in the context of a project aiming at the identification of cfDNA biomarkers for their diagnostic utilization via quantitative real-time PCR (qPCR) based assays. The concept of this work resides on three main pillars, each comprising an in individual objective. The first is the already mentioned search for diagnostic cfDNA biomarkers. It involves the targeted analysis of large amounts of sequencing data where it is tried to correlate the increased or decreased occurrence of certain sequence motifs to a distinct disease condition. The second is the investigation of this large amount of data in order to obtain new insights into the biology and properties of cfDNA, where its composition was analyzed in detail for cues on eventual mechanisms of origin. The third is the comparison and alignment of all our observations and findings with the scientific literature with the goal of creating a combined image. Each of these pillars comprise an individual chapter of this thesis and a research article. These are introduced in more detail within the prologue and article introduction of the respective chapter.

Chapter 2

Analysis of cell-free DNA sequencing data

2.1 Objectives

My approach towards the analysis of whole-genome cfDNA sequencing data relied on two main hypotheses made by our research group:

- The coverage of cfDNA sequencing data is not uniformly distributed when mapped against the genome. This implies that there are genomic regions with comparatively high coverage in contrast to other regions being either weakly covered or uncovered. In contrast, sequencing genomic nuclear DNA, *i.e.* from disrupted cells of a tissue culture, would be expected to result in an approximately even/uniform distribution of coverage, provided that the genome model is accurate.
- A host's physiological condition may reflect in the under/over representation of certain sequence motifs.

The primary goal of our research was to identify genomic regions featuring high cfDNA coverage in correlation to specific host conditions and explore their usability as biomarkers for diagnosis via qPCR. A main objective of my work could therefore be formulated as design and development of a procedure for the quantitative analysis and comparison of read coverages from two groups of samples (*e.g* disease and control). In addition to detecting differentially covered genomic regions for their assessment as diagnostic biomarkers, a second goal was to obtain insights on potential mechanisms behind the presence of predominantly occurring sequence elements by associating them with annotations.
Similar requirements regarding quantitative evaluation of sequencing coverage are also encountered in techniques like RNA Sequencing Differential Expression (RNA-Seq DE) [44, 45] or DNA enrichment methods, such as Methylated DNA Immuno-Precipitation (MeDIP). MeDIP is a technique for the identification and quantification of methylated DNA using 5-methylcytosine specific antibodies where the immuno-precipitated DNA can for example be sequenced [46, 47]. RNA-Seq DE aims at the identification of differently expressed genes between samples in connection to certain biological conditions. This technique involves isolation and purification of steady-state RNA, which is converted to cDNA via reverse transcription before sequencing. The identification of significant changes in expression levels requires assigning the sequenced reads to the exons of genes under consideration of mRNA splicing, transcript variants and reads from exon boundaries or introns. The obtained read counts have to be subjected to normalization strategies, which account for sequencing library sizes as well as for transcript lengths [44].

Numerous tools have been developed, which either use existing gene annotations or perform de-novo transcriptome assembly and apply different statistical models and normalization techniques in order to assess the significance of resulting fold changes [45].
To my knowledge, all Bioinformatics tools in this context operate on the level of read counts which are assigned to clearly defined genomic regions. Their definition is either based on existing annotations or occurs via tiling/binning strategies, where the genome is divided into adjacent windows of a predefined size [45, 47]. Since no prior knowledge concerning the relevance of distinct genomic elements in context with cfDNA existed, a potential approach could not rely on existing annotations. Arbitrary tiling of the genome comes with the disadvantages that information about exact read positions gets lost and certain regions may be cut-off or split in half. Therefore, a strategy which analyses the genome in nucleotide resolution and does not rely on partitioning was desirable. This led to the decision of developing a dedicated software for the processing and analysis of coverage data based on genomic alignments created from raw data (see section 2.3). While the development of software always is a risk under consideration of a limited project time frame, it comes with the advantage of allowing a tailored analysis instead of being restricted to the functionality of existing tools. Having full access to all routines, adaptation and extension of a framework is easy and does not require understanding and editing foreign source code. Alternatively, an analysis procedure can be built via a combination of existing software solutions in the form of a so-called Bioinformatics pipeline [48, 49]. The individual steps of an analysis often require the application of multiple tools, wherefore pipelines tend to become very complex. This complexity increases the probability for errors and causes problems due to versional changes and dependencies [34]. Another serious drawback of large pipelines is the consecutive generation of intermediate data, where each program writes its output to the filesystem in turn being loaded again by the next program. This can substantially slow down execution, especially when the access to mass storage occurs via network. Thus, the requirement for the analysis of hundreds of samples, comprising terabytes of data, was a major motivation behind the development of a dedicated analysis framework, where the ideas and theory behind its design are described by this chapter.

2.2 Properties of the data

My approach to analyze the distribution of cfDNA coverage over the genome in nucleotide resolution required mapping of the sequencing data to the respective genome and subsequent translation into coverage data. A schematic overview of this process is provided by Figure 2.2. The following subsections introduce the different forms of data encountered in this process, and discuss their origin, properties and considerations regarding efficient computational processing and analysis.

2.2.1 Paired-end DNA sequencing data

Paired-end next generation sequencing is nowadays mostly performed using the Illumina sequencing by synthesis chemistry [50, 51], which is based on the principles of Sanger's and Coulson's chain terminator method [18]. Before the actual sequencing process 3′ and 5′ adapters have to be added to the target DNA via limited-cycle PCR [52]. It then is diluted to a fixed concentration before being loaded into the flow cell, where a lawn of complementary adapters is covalently bound to the glass surface at their 5′ end. Individual DNA fragments bind to adapters and are amplified by bridge PCR [50], where alternately forward and reverse complement copies of the original fragment are generated until a clonal cluster is formed. A fixed DNA concentration before loading the flow cell is necessary to avoid overlapping clusters or underpopulated flow cells. All clusters within the flow cell are sequenced simultaneously, where at first a primer is used which binds the forward copies for sequencing all clusters from one end. Primers are elongated by polymerase, incorporating fluorescently labeled 3′-O-azidomethyl 2′-deoxynucleoside triphosphates as reversible terminators [50] and base calls are made via fluorophore excitation. Subsequently the flow cell is flushed and terminators plus fluorescent dye are removed. The process continues repetitively until the sequence of the so-called "first read" is complete for each cluster. Ultimately, the sequencing process is performed a second time using a primer for the reverse complement copies, where the "second read", being also called "mate", is obtained for the respective clusters. Thus, each fragment is practically sequenced from both ends, resulting in paired-end sequencing data [53, 54]. This technology does not suffer from erroneous base calling due to homo-polymeric sequences, but the signal degrades with increasing read length. This is because usually not 100% of all copies within a clonal cluster are elongated in each cycle, leaving a fraction of fragments where the synthesis is lagging behind whereby the signals blurs out. The sequencing data is usually stored in FastQ [55] formatted text files, containing the sequences of all reads plus a quality score for each base, which is derived from the signal strength of the respective base call. Raw sequences often contain parts of sequencing adapters [56, 57]. Removal of those remnants and trimming of low quality read ends [58] via specialized tools like Cutadapt or Trimmomatic [56, 57] is recommended before processing of the data.

The strategy for producing sequencing data for our project aimed on the isolation of lowly concentrated cfDNA from human and animal blood serum or plasma. For this purpose, the High Pure Viral Nucleic Acid Kit (Roche Applied Sciences 11858874001) was used by our lab team, which is designed for isolating low concentrations of viral DNA from whole blood, serum or plasma. It uses a lysis buffer and proteinase K to free DNA fragments from proteins and lipid membrane residues. The extraction of nucleic acids occurs via specific binding to the surface of silicate fibers in presence of a chaotropic salt. DNA isolation was followed by amplification and purification (for details see chapters 4 and 5). Library preparation and Illumina paired-end [53, 54] whole-genome DNA sequencing was performed by SEQ-IT GmbH, Kaiserslautern, Germany on a NextSeq™ 500 platform, which is advertised with a Phred quality score [59, 60] Q30 (99.9% accuracy, 1 error in 1000 base calls) for over 75 % of the base calls. The DNA sequencing libraries were generated by SEQ-IT with the Nextera™ XT DNA Sample Preparation Kit [61] (Illumina, FC-131-1096). In this process, the DNA was fragmented and covalently tagged simultaneously by in vitro transposition of engineered Nextera™ Transposome complexes with free transposon-DNA ends. This single-step library preparation approach avoids sample loss and is therefore well suited for preparing samples with small amounts of DNA [52, 61].

2.2.2 Genomic Alignments

Genomic alignments are mappings of sequencing data to an organism's genome, which is referred to as template. They are *e.g.* required for the evaluation of coverages or read-depths at genomic regions and are calculated via specialized highly

efficient search algorithms [62]. Due to inevitable contaminations with foreign DNA, sequencing errors, imperfect templates, DNA-amplification errors and natural genetic variations an exact match for a distinct read may not always exist [63]. Therefore, search algorithms have to allow for inexact or gapped read alignments. Scoring the degree of agreement with the template is necessary for identifying the best matching positions. A scoring threshold is used for discarding reads as unmapped given that no sufficiently similar sequence on the template is found. In cases where more than one equally scored locus exists either all loci may be reported or one will be chosen stochastically. In the context of quantitative analysis of coverage data it is important that the latter procedure is used. Otherwise multi-copy and repeat regions will automatically receive increased coverage levels. Paired-end sequencing data provides additional information for identifying the most probable location of a read pair. In this case both reads stem from the same DNA fragment, which is supposed to be sized within in a certain range of length depending on the fragmentation step during library preparation. Thus, both reads are expected to align in a certain orientation and approximate distance to each other. Such an alignment is referred to as "concordantly mapped" read pair. Finding the most likely locus of origin for large number of reads is a computationally costly problem under consideration of the stated properties [64] and the large search space comprised by genomes of higher organisms [65].
Genomic alignments are stored in Sequence/Binary Alignment Map (SAM/BAM) [66] formatted files. While SAM files are text files, a BAM file is a binary representation of a SAM file, where zlib-compression is applied for reduced file sizes. The tab-delimited, text based SAM format was designed to store different types of biological sequence data (*e.g* DNA, RNA or AAs). It is structured into a header section, where amongst other information mainly the properties (primarily name and length) of reference sequences are declared, and an alignment section. The sequence data in the latter is organized line-wise and may either be aligned or unaligned. Each line consists of an sequence identifier and a set of fields, where information about the sequence and the alignment (*e.g* reference name, position) are stored. SAM/BAM files can also be sorted by coordinates and indexed in order to facilitate efficient search and data processing. More detailed information may be looked up in the format specification [67].

The alignment program used in this study was Bowtie 2 [68]. It uses a precomputed search index of the template for improved mapping speed, which is a trade-off in terms of memory usage [62]. The indexing strategy used by Bowtie 2 is an extension of the full-text minute index [69], which was adapted to allow for gapped alignments.

2.2.3 Count- and coverage data

The information how many reads are mapped to distinct regions of the genome is referred to as count data or coverage. In order to represent coverage data in nucleotide resolution, it is necessary to store an absolute or relative coverage value for every nucleotide position on the genome. Absolute coverage values specify the exact number of mapped reads at a distinct nucleotide position. Relative values are the result of a normalization strategy and depict the coverage in relation to a reference measure. At the level of coverage data sequences of individual reads are no more of central importance, but may be efficiently retrieved from indexed sorted alignment files whenever necessary.

Fragment reconstruction

Since both reads of a read pair in paired-end sequencing data stem from a single DNA fragment, additional information may be inferred during the generation of count data. Two scenarios have to be considered for concordantly aligned read-pairs. Both reads may either align apart of each other or the alignment might overlap. Processing the reads of an overlapping pair individually would produce double-counting of the positions within the overlap, whereas originally only a single DNA fragment was present. Given that both reads align with some distance to each other, it may be presumed that the nucleotides within the gap still existed on the original DNA fragment. Therefore, concordantly aligned read pairs can be treated as one single DNA fragment during coverage data generation (see section 2.3.1). By incrementing the count values from the start of the first read to the end of the second read of a pair, double counting at overlaps is avoided and additional count information is obtained from non-overlapping alignments (see Figure 2.1). This approach differs from classical end-joining or merging techniques [70] in a major aspect. Fragment reconstruction occurs after aligning the read pairs to the reference genome. This utilizes additional information from the template and is less error prone than simply matching an eventual overlap of a read pair, which is significantly impaired by sequencing errors due to quality loss at read ends [70].

Normalization

Even if all samples in an analysis are ordered with the same sequencing depth from a sequencing company, usually there exist aberrations in read quantities between samples in a scale around 10 %. Also, the yield of successfully mapped reads or read pairs during alignment varies due to small inconsistencies in sample properties, DNA preparation, inevitable contaminations and other influences. Applying normalization to count data provides invariance between samples and improves the quality of an analysis. Several methods for the normalization of sequencing data exist, being mostly applied in the field of RNASeq DE, where gene expression levels are compared on the basis of reads mapped to exonic regions [71]. For example, the Reads/Fragments Per Kilobase per Million mapped reads/fragments (RPKM/FPKM) method or scaling of gene expression levels in regard to reference housekeeping genes are frequently used for data normalization [71]. In an annotation free approach these methods cannot be applied. In such a case, a total count normalization approach, where all values are normalized in regard to the total amount of mapped reads or nucleotides, can be used for the normalization of coverage data.

2.3 Preprocessing: Efficient access to the data

Fast and efficient access to coverage data was the foundation to all analyses described within this chapter. They shared the same data preprocessing procedure, as illustrated in Figure 2.2, involving quality and adapter trimming, quality control, mapping and calculation of coverage data. The initial steps for the creation of BAM formatted genomic alignments were performed with established Bioinformatics tools. FastQC [72] and MultiQC [73] were used to check read qualities and adapter content before and after adapter removal with Trimmomatic [57]. Bowtie 2 [68] and SAMtools [66] were used for the creation of alignments and for the conversion of SAM files to sorted BAM files. The additional I/O time consumption caused by repeated writing and loading of intermediate data through the execution of these tools was negligible in comparison to the generation of genomic alignments and had to be performed only once for a given dataset.

BAM formatted alignments were the entry point for the Java™ framework implemented in the context of this work. It included the functionality for the calculation of coverage data and all subsequent analyses and evaluations. The framework was designed to deal with terabytes of data from hundreds of samples. Loading the total data of one or more datasets of such proportions would usually exceed the memory of most computers. Under these considerations, it is advantageous if analyses can be performed chromosome-wise, loading samples one by one for calculations. This strategy was pursued for implementing memory consuming procedures while also focusing on fast data access (see section 2.3.2).

2.3.1 Calculation of coverage data

In order to obtain coverage data, all records of a BAM formatted alignment were parsed using the Picard Java™ library [74]. Start and end coordinates from aligned reads were extracted and sorted according to chromosomes, where concordantly mapped reads were stored separately and joined with their mates. The absolute coverage data $cov_{Ai}(c, p)$ of a sample i was then calculated chromosome by chromosome, where c is the chromosome number and p refers to a nucleotide position on the respective chromosome. It was stored in an integer array, and created by incrementing the values for every corresponding position from the start of the first read to the end of the second read for each concordantly mapped read pair (see 2.2.3). Optionally, the coverage of discordantly mapped reads could be added as well, although this was omitted in our analyses since discordant mappings are less reliable. The normalized relative coverage data $cov_{Ri}(c, p)$ was subsequently calculated by dividing every position p through the average coverage of the sample, which is the sum $\sum_{c,p} cov_{Ai}(c, p)$ divided by the size of the genome. This comprised a form of total count normalization scaled by the genome size, where values over 1 represented above average coverage. Chromosome sized integer and float arrays for absolute and relative coverage date were finally LZ4 [55] compressed and written to files on mass storage.

2.3.2 Storing and compression

While the size of alignment data was mainly dependent on length and number of mapped reads, the size of coverage data in nucleotide resolution depended plainly on the reference genome size. Alignments produced from 10 million reads of 150 bp had a file size of about 5 GB in SAM format or 1 GB in BAM format. The representation of the corresponding coverage data for a mammalian genome of 3 Gbp (human genome 3.29 Gbb, bovine 2.72 Gbp) required about 12 GB when using a 32 bit integer or float representation for every nucleotide position. Because this type of data was purely numeric, it could be efficiently compressed to only a few megabytes using *i.e.* GZIP compression. This did not only save storage space, but also accelerated loading of data into the memory significantly. Since the data needed to be decompressed, this comprised a trade-off between computation time and I/O. LZ4 [55] compression provides extremely fast compression and decompression at the cost of reduced compression rates, wherefore it was used as a mean for fast and efficient data access in this framework. The largest human chromosome consists of 247 million bp, which required about 3 MB of storage when GZIP compressed and about 6 MB when LZ4 compression was applied.

2.4 Analysis of cfDNA coverage

The relative coverage cov_{Ri} of a sample i can be viewed as complex probability distributions over all nucleotide positions of the genome, where $cov_{Ri}(c,p)$ is proportional to the probability that a sequenced read will be mapped to the specific genomic position. Sampling, or rather sequencing, a finite number of reads from the sample yields a noisy representation of the true coverage distribution of the respective sample. In case of average count normalization this probability distribution is simply scaled by the size of genome. The coverage distribution of a sample set S (*e.g.* disease or control) was represented as an array of univariate Gaussian distributions given by

$$COV_S(c,p) = \mathcal{N}_S(\mu(c,p), \sigma^2(c,p))$$

with an arithmetic mean $\mu(c,p)$ and a variance $\sigma^2(c,p)$ for every genomic nucleotide position. These parameters were computed from the normalized coverages of all samples in the set. To achieve reasonable memory usage, the computation was performed separately for each chromosome by sequential loading and decompression of the individual relative sample coverages (see 2.3.1). The process involved additions of very large arrays, which was sped up by parallelization, where the scope of indices was partitioned and individual threads performed the additions within a distinct section. The implementation used multi-threading for summing coverage values and squared deviations during the computation of mean and variance. The resulting coverage distributions were used for the prediction of biomarkers and also provided the basis for the analysis of the properties of cfDNA coverage (see Chapter 4). Coverage distributions could also be computed for specific genomic loci, where only coverage data from the respective chromosome was loaded, and computations were only performed for the corresponding positions. This was *e.g.* used for the generation of coverage plots, as depicted in Figure 2.3.

2.5 Identification of cfDNA biomarkers

A candidate biomarker for a distinct physiological condition of subjects was a genomic region of a certain size, where the cfDNA coverage differed significantly between subjects featuring the specific condition and a healthy control group (see Figure 2.3. In order to be suitable for PCR experiments, the regions needed to have a length of 100 to 400 bp and to allow the design of primers which align to exactly one location at the respective genome. Determination of such regions required the computation and comparison of the coverage distributions from both groups, as illustrated in Figure 2.4. A commonly used measure for comparing probability distributions [75] is the Kullback-Leibler Divergence KL [76] given by

$$KL(P,Q) = \int_{-\infty}^{\infty} p(x) \cdot \log \frac{p(x)}{q(x)} \cdot \mathrm{d}x,$$

where P and Q refer to arbitrary probability distributions with probability density functions $p(x)$ and $q(x)$. In case of two univariate Gaussians, the Kullback-Leibler Divergence is given by

$$KL(\mathcal{N}_1, \mathcal{N}_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

A drawback of this measure is, that it is not symmetric. Thus the comparison of P with Q yields a different result than comparing Q with P.

$$KL(P,Q) \neq KL(Q,P)$$

It can be made symmetric easily by combining both comparisons.

$$KL(P,Q) + KL(Q,P)$$

For univariate Gaussian distributions this combined measure is given by

$$= \frac{1}{2} \cdot \left(\frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_2^2 + (\mu_2 - \mu_1)^2}{\sigma_1^2} \right) - 1,$$

where the cancellation of the logarithmic terms is beneficial since log operations are computationally costly. Adding a small offset value ϵ to both variances is necessary to ensure that the KL is always defined.

$$= \frac{1}{2} \cdot \left(\frac{(\sigma_1 + \epsilon)^2 + (\mu_1 - \mu_2)^2}{(\sigma_2 + \epsilon)^2} + \frac{(\sigma_2 + \epsilon)^2 + (\mu_2 - \mu_1)^2}{(\sigma_1 + \epsilon)^2} \right) - 1$$

2.5.1 Sliding window scan and de-noising

The search for candidate marker regions with a predefined length l was implemented as sliding window scan and was also performed chromosome wise for reduced memory usage. A sliding window of size l propagated over all nucleotide positions of each chromosome, where the sum of values within the window divided by l was computed, yielding an average window coverage. The windowed relative coverage of a sample i on window position p^W is given by

$$cov_{Ri}^{W}(c, p^W) = \frac{\sum_{p=p^W}^{p^W+l-1} cov_{Ri}(c, p)}{l}.$$

During window propagation this sum was efficiently updated for each step by adding newly entering values and subtracting values, which left the window. Perpetual addition and subtraction is prone for accumulating numerical precision errors, wherefore the sum was computed in double precision instead of float precision. Genome models often contain regions, where the exact nucleotide sequence is assumed to be unknown, which is denoted by the character N instead of the nucleotide representing characters A,T,G and C. Window positions containing Ns were excluded from the analysis by masking. Computation of an average window coverage for every position came with the advantage of a smoothing effect which reduced the impact of individual nucleotide positions and facilitated a more robust estimation of coverage distributions, where the windowed coverage distribution of a sample set S is denoted by $COV_S^W(c, p^W)$.

The results of our analysis showed a rather non-uniform distribution of cfDNA coverage over the genome (see Chapters 4 and 5), featuring strongly covered regions interlaced with practically uncovered or weakly covered areas. Weak coverage could be either attributed to very low abundance of the respective DNA fragments in the sample or to noise introduced during sample preparation, sequencing and data processing, as outlined in section 2.2. Either way, the low number of mapped reads at such regions entailed an unstable estimation of coverage distributions, which as a consequence impaired the detection of reliable marker candidates. In order to overcome this issue, the windowed coverage was subjected to an optional noise filtering via application of a sigmoid function

$$f(x) = \frac{1}{1 + e^{-(sx+b)}},$$

where the parameters s and b could be adjusted for different levels of noise reduction. Weakly covered regions in the coverage distribution of a sample could also occur due to regions which were strongly covered in only one or a few samples, while being uncovered or weakly covered in the remaining samples of the respective set. Strong coverage at individual samples was not supposed to be considered as noise, although it still could be filtered during the optional outlier removal procedure. Therefore, the sliding window scan and noise filtering occurred at the level of individual samples rather than at the level of coverage distributions. Since the sliding window scan had to be performed multiple times for each sample in order to estimate windowed coverage distributions, it constituted a performance critical procedure. It was parallelized using the same index partitioning strategy that was used for array additions, where an additional thread was used for prefetching and decompression of the next sample during execution.

2.5.2 Outlier removal

Regions with extreme coverage values may sometimes occur at distinct samples, whether being caused by PCR or sequencing artefacts, or being genuine features of the sample, is to be left unanswered. Such regions may be considered as outliers which significantly distort the estimation of coverage distributions and can be filtered in an optional outlier removal procedure.

A simple technique to identify and remove outliers is z-scoring [77, 78]. In case of univariate data the z-score z_i is computed for a data point x_i as the absolute difference from an initially computed arithmetic mean $\bar{x}$ divided by the standard deviation s.

$$z_i = \frac{|x_i - \bar{x}|}{s}$$

Based on a threshold, *e.g.* $z_i > 2.5$, it is then decided if x_i is considered as outlier or not. The statistics of the data are then computed again under exclusion of data points flagged as outlier.

For windowed coverage distributions outliers could only be removed position wise. Otherwise, the procedure would have led to the removal of every sample featuring an extraordinarily covered region at some genomic location. After initial estimation of a windowed coverage distribution, the sliding window scan was performed again, where for every position of the sliding window p^W the z-score given by

$$z_i(c, p^W) = \frac{|cov_{Ri}^W(c, p^W) - \mu^W(c, p^W)|}{\sigma^W(c, p^W)}$$

was determined. A new windowed coverage distribution was then computed, omitting samples being flagged as outlier only at distinct window positions.

2.5.3 Determination of candidate marker regions

After the windowed coverage distributions $COV^{W}_{Disease}(c, p^{W})$ and $COV^{W}_{control}(c, p^{W})$ were determined, the symmetric, combined KL (see section 2.5) was computed, measuring the difference of window coverages between both groups. The resulting sequence of values were subsequently searched for peaks, where a preset number of window positions with the largest KL values was reported. A distinct peak could either occur as individual peak value, where the positions before and after were smaller, or as sequence of identical values with the same property. In the latter case the middle of such a plateau was reported. Peaks could be very close to each other. Neighborhood suppression was performed in order to avoid reporting the same genomic region, set apart by only a few bases, multiple times. Hereby the largest peak suppressed all other peaks within a neighborhood of a certain size.

2.5.4 Candidate marker evaluation

Regions, differing in coverage distribution between disease and healthy sample sets without correlation to the physiological condition of interest, were expected to occur due to pure coincidence considering the very large search space comprised by a mammalian genome. In order to improve the prediction of suitable marker regions, a pre-evaluation of their discriminative power was performed. Assuming that a marker region systematically featured stronger coverage for one of both conditions, it was searched for an optimal threshold value separating samples from both classes. This was done by simply sorting the samples by coverage values at a distinct marker region and by finding an ideal bipartition of the sort sequence. The resulting discriminative performance specified the percentage of correctly assigned samples to the respective conditions based on coverage data. In case of imbalanced sample sets correspondingly scaled weights were applied so that both sets contributed equally to this performance score. Since markers were supposed to be evaluated relatively to each other in qPCR experiments (see chapters 3 and 5), a pair-wise performance pre-evaluation was also performed. Hereby, the ratios or difference between coverages of two markers within the subset were used for another threshold-based discrimination of samples. An ideal subset of candidates was ultimately selected from the best performing markers or marker pairs and a detailed report was generated for these final candidates. Amongst other information, such as sequence similarities between marker candidates, the report contained genomic coordinates and coverage plots for the marker regions and their surroundings as well as template sequences and consensus sequences generated from the sequencing data aligned to the respective marker regions. The latter showed the color-coded agreement between the reads mapped to these regions for both the sample sets, which is used for PCR primer design.

2.6 Analysis of cfDNA composition and properties

In order to compute the composition of cfDNA sequencing data in regard to certain annotated genomic elements, the coverage data of a sample had to be assigned to the respective elements. We *e.g* analyzed the composition of cfDNA in terms of repeat families annotated in the UCSC Repeat Masker database [79] as described in Chapter 4. A very similar analysis was performed by Bronkhorst *et al.* 2016 [80] for cfDNA from cancer cell culture supernatants, where they assigned the reads to annotation elements using BLAST [31].

The fast access to coverage data provided by the framework described in this chapter also allowed very exact and efficient calculations of compositions in a straightforward manner, which was again performed chromosome-wise for reduced memory usage. Since a composition is inherently normalized, it can be directly computed from the absolute coverage $cov_{Ai}(c, p)$ of a sample i. For this, all annotation records of a certain repeat family were read out from the database. They were subsequently translated into a binary mask marking all genomic areas where the respective repeat family was annotated. These binary masks were also compressed and written to mass storage for future usage, avoiding repetitive generation. Using masks for all considered repeat families, the coverage of a sample was assigned to repeat families or to non-repetitive areas. In case of overlapping repeat annotations, the coverage was split equally between the respective repeat families. Once the coverage of all genomic positions was assigned, the shares of every repeat family were divided by the total sample coverage, yielding their percentage of the composition. This percentage is also referred to as relative coverage of a repeat family given by

$$cov_{Ri}(Rfam) = 100 \cdot \frac{\sum_{\forall c,p \in P_{Rfam}} \frac{cov_{Ai}(c,p)}{n_{Rfam}(c,p)}}{\sum_{\forall c,p} cov_{Ai}(c,p)},$$

where P_{Rfam} is the set of all genomic positions where repeat family $Rfam$ is annotated and n_{Rfam} refers to the number of repeat families annotated at a specific position.

In order to determine if a repeats family fraction was over- or under-represented within a composition, it had to be put in relation to the genomic abundance of the repeat family. Accounting for overlapping annotations, the abundance of a repeat family was calculated analogously and is given by

$$abundance(Rfam) = 100 \cdot \frac{\sum_{\forall c,p \in P_{Rfam}} \frac{1}{n_{Rfam}(c,p)}}{\sum_{\forall c,p} 1}.$$

The relation $\frac{cov_{Ri}(Rfam)}{abundance(Rfam)}$ is being referred to as representation of the repeat family, where values above or below 100% correspond to over- or under-representation, respectively.

The statistical analysis performed on cfDNA compositions are described in Chapter 4.

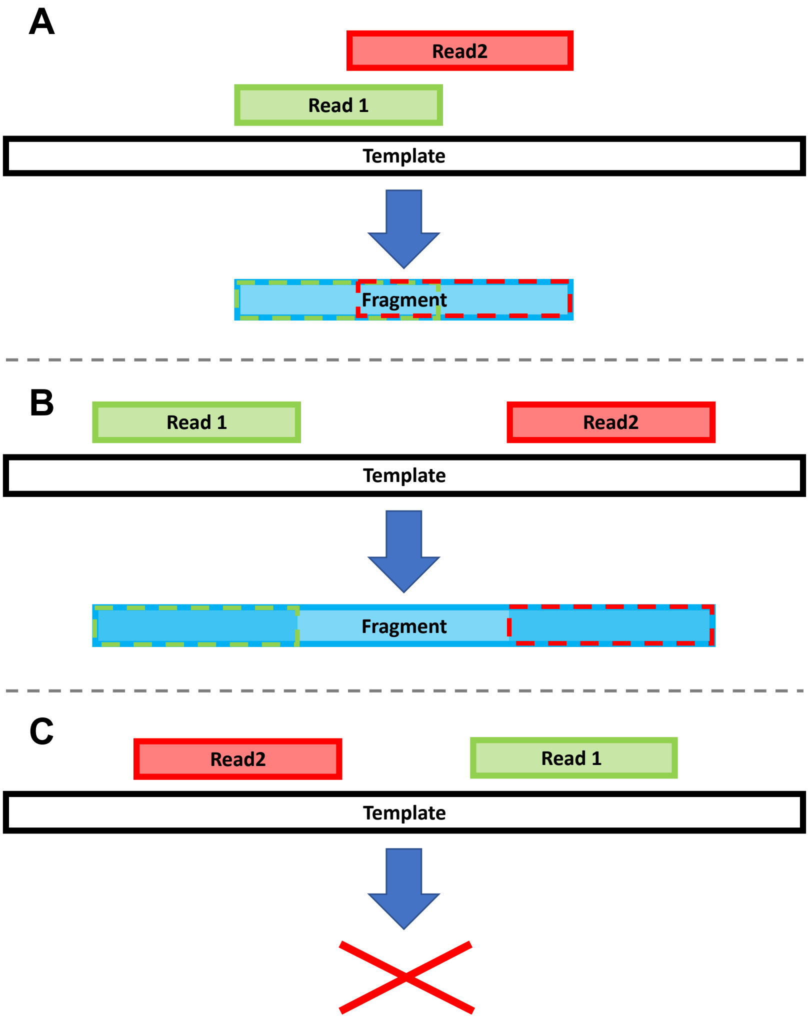


FIGURE 2.1: Illustration of fragment reconstruction. The green and red rectangles signify the first read and the respective mate read of a read pair. When mapped concordantly, either overlapping (A) or with a certain distance in-between (B), these reads can be interpreted as a single DNA fragment, indicated by the blue rectangles. When mapped discordantly, *i.e.* in the wrong orientation to each other (C) or far apart, no fragment can be reconstructed.

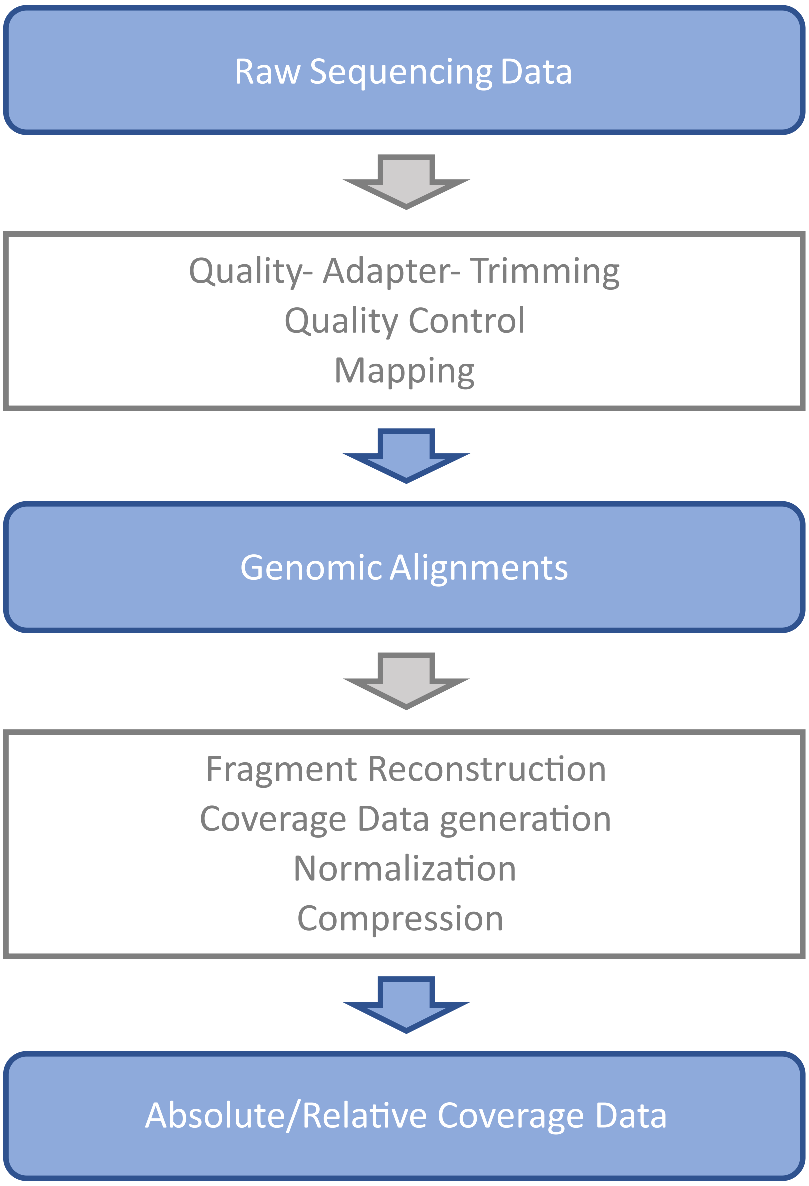


FIGURE 2.2: Schematic overview of the coverage data generation from the raw sequencing data. In succession to adapter removal and quality control the sequencing data was mapped to a reference genome. The resulting genomic alignments served as the basis for the reconstruction of fragments from concordantly mapped read pairs and for the calculation of coverage data. A detailed description of this preprocessing phase can be found in section 2.3.

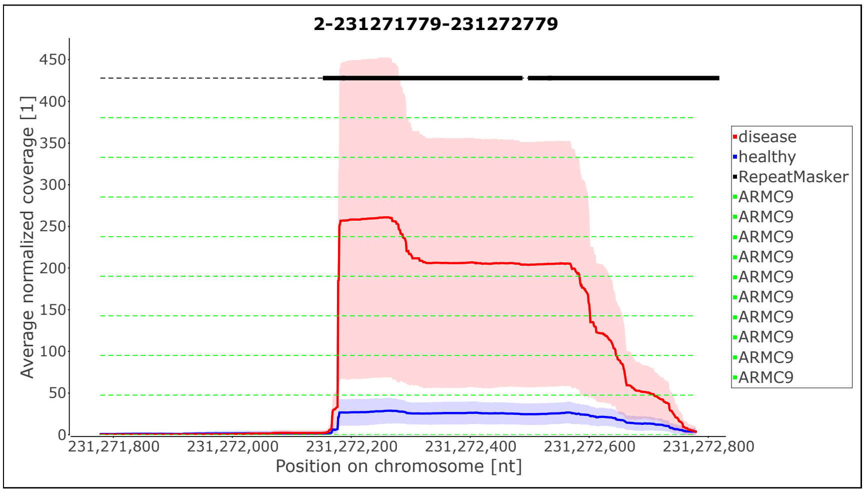


FIGURE 2.3: Example of a coverage plot. The genomic locus features a region on chromosome 2 with significantly increased coverage in a sample set featuring a disease condition (red) when compared to a healthy control group (blue). The solid blue and red lines signify the mean coverage of the respective sample and the shaded areas around them signify the standard deviation. The coverage peak is located on two adjacent repeat regions, as indicated by the black bars on the top, on an intron of the *ARM9* gene, as indicated by the green dashed line. Since multiple transcript variants of this gene are annotated, there are multiple green dashed lines.

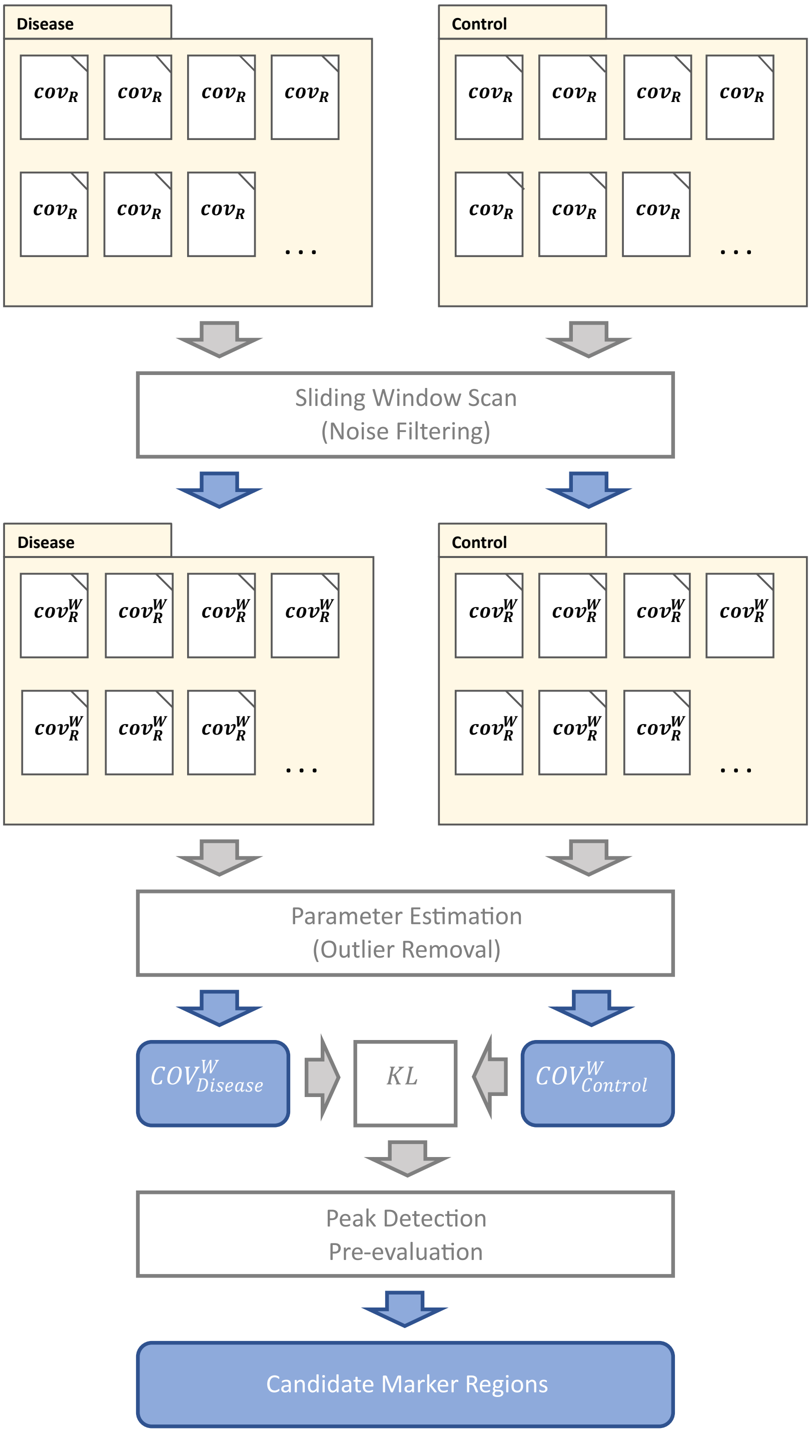


FIGURE 2.4: Schematic overview of the marker prediction procedure. A sliding window scan with optional noise filtering was applied to every sample in the sample sets *Disease* and *Control*. The coverage distributions for both sets were subsequently estimated from all samples of a set, where an optional outlier removal could be performed. Differences between both sets were quantified using the Kullback-Leibler Divergence KL. Candidate marker regions were identified via peak detection and were subjected to a pre-evaluation of their estimated discriminative performance.

Chapter 3

Overview of the biology and origins of cell-free DNA in humans

3.1 Prologue

The review article introduced in this chapter was the third and final publication of my PhD time. Tough being the last work, it's enlisted first in this thesis, because it reviews the current state of knowledge on the biological backgrounds of cfDNA and therefore is a good entry point for a reader. At the time, when our article about cfDNA biomarkers for the detection of human sepsis (see Chapter 5) was in the peer reviewing process, we began to plan a review article which aligns our findings with the scientific literature and creates a combined picture. The second article about the composition of cfDNA regarding different repeat families (see Chapter 4) was freshly submitted, when we discovered an article from Bronkorst *et al.* 2016 [80], who investigated cfDNA from cancer cell-cultures and reported findings which were very similar to ours. We decided to contact Dr. Abel J. Bronkhorst and his fellow researchers and expressed our interest to work on a review article together. I suggested to focus the scope of this review article primarily on the biology and experimental findings behind the various putative origins of cell free DNA, since most of the literature addressed this topic mainly from the perspective of clinical diagnostics of diseases. At this time, I already had composed a first draft about extracellular vesicle associated cfDNA, which I proposed as an example how a chapter of this article could look like. The idea was received positively, and we decided to divide the work on different biological release mechanisms of cfDNA amongst us authors, where each mechanism was supposed to be addressed in an individual chapter. Thus, Abel designed a chapter on chromosomal instabilities and the extrusion of micronuclei, while Dr. Klaus Schliep devised a chapter on the secretion of nucleoprotein complexes and Dr. Ingund Rosales Rodriguez and Daniel Schwendenwein addressed cell death and extracellular traps. In addition to the chapter on vesicle associated cfDNA I contributed a second chapter about erythroblast enucleation before we combined everything together into a consistent piece of text, which was complemented with illustrative figures created by Vida Ungerer.

3.2 Article

The article was accepted for publication in the MDPI International Journal of Molecular Sciences on October 27th 2020 and published on October 29th 2020.
DOI: 10.3390/ijms21218062

3.3 Author rights and permissions

Articles published in MDPI journals are licensed under the Creative Commons CC BY 4.0 license, where copyright is retained by the authors. Reuse and quotation is permitted provided that the original publication is properly cited.
https://www.mdpi.com/authors/rights
https://creativecommons.org/licenses/by/4.0/

International Journal of
Molecular Sciences

MDPI

Review

Putative Origins of Cell-Free DNA in Humans: A Review of Active and Passive Nucleic Acid Release Mechanisms

Stefan Grabuschnig [1,†], Abel Jacobus Bronkhorst [2,†], Stefan Holdenrieder [2], Ingund Rosales Rodriguez [3], Klaus Peter Schliep [1], Daniel Schwendenwein [3], Vida Ungerer [2] and Christoph Wilhelm Sensen [1,3,4,*]

[1] Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14(V), 8010 Graz, Austria; stefan.grabuschnig@tugraz.at (S.G.); klaus.schliep@tugraz.at (K.P.S.)
[2] Institute for Laboratory Medicine, German Heart Centre, Technical University Munich, Lazarettstraße 36, 80636 Munich, Germany; bronkhorst@dhm.mhn.de (A.J.B.); holdenrieder@dhm.mhn.de (S.H.); ungerer@dhm.mhn.de (V.U.)
[3] CNA Diagnostics GmbH, Parkring 18, 8074 Grambach, Austria; ingund@cnadiagnostics.com (I.R.R.); daniel@cnadiagnostics.com (D.S.)
[4] BioTechMed Graz, Mozartgasse 12/II, 8010 Graz, Austria
* Correspondence: csensen@tugraz.at; Tel.: +43-664-608734090
† These authors contributed equally.

Received: 16 October 2020; Accepted: 27 October 2020; Published: 29 October 2020

Abstract: Through various pathways of cell death, degradation, and regulated extrusion, partial or complete genomes of various origins (e.g., host cells, fetal cells, and infiltrating viruses and microbes) are continuously shed into human body fluids in the form of segmented cell-free DNA (cfDNA) molecules. While the genetic complexity of total cfDNA is vast, the development of progressively efficient extraction, high-throughput sequencing, characterization via bioinformatics procedures, and detection have resulted in increasingly accurate partitioning and profiling of cfDNA subtypes. Not surprisingly, cfDNA analysis is emerging as a powerful clinical tool in many branches of medicine. In addition, the low invasiveness of longitudinal cfDNA sampling provides unprecedented access to study temporal genomic changes in a variety of contexts. However, the genetic diversity of cfDNA is also a great source of ambiguity and poses significant experimental and analytical challenges. For example, the cfDNA population in the bloodstream is heterogeneous and also fluctuates dynamically, differs between individuals, and exhibits numerous overlapping features despite often originating from different sources and processes. Therefore, a deeper understanding of the determining variables that impact the properties of cfDNA is crucial, however, thus far, is largely lacking. In this work we review recent and historical research on active vs. passive release mechanisms and estimate the significance and extent of their contribution to the composition of cfDNA.

Keywords: cell-free DNA; circulating DNA; liquid biopsy; circulating tumor DNA; active release of cfDNA; passive release of cfDNA; origins of cfDNA

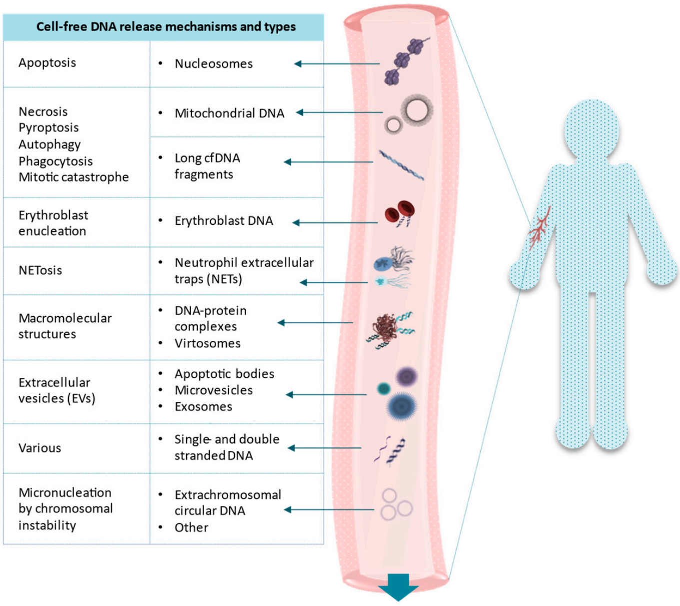
1. Introduction

Cell-free DNA (cfDNA) is a fraction of circulating nucleic acids (CNAs), which was discovered and first described by Mandel and Metais, in 1948 [1]. The term encompasses all kinds of extracellular DNA molecules found in serum or plasma and other body fluids [2] of vertebrates and includes genomic and mitochondrial host DNA [3,4], as well as foreign DNA [5,6], for example, of bacterial or viral origin. The cfDNA molecules occur predominantly in the form of double-stranded DNA (dsDNA) [7] and are mostly of small size, ranging between 100 and 200 base pairs (bp) [8]. Larger fragments, even in the range of several kilobase pairs (Kbp) have also been reported [9–11].

In 1973, Leon et al. showed that cancer patients exhibited elevated levels of cfDNA in serum [12]. For this purpose, the group developed and performed a radioimmunoassay, where ^{125}I-Iododeoxyuridine labelled DNA was detected via antibodies from a lupus erythematosus patient. Similar results had already been described by Tan et al., in 1966, [13] for systemic lupus erythematosus, and even as early as 1948, Mandel and Metais [1] showed that cfDNA levels were elevated in several disease conditions. Although the cfDNA concentrations in the serum declined in correlation with improved clinical conditions after treatment, Leon et al. were initially skeptical about the potential diagnostic value of cfDNA, because approximately half of the cancer patients had serum cfDNA concentrations that were within the same low range as found in healthy subjects. However, the group kept pursuing the idea of a cfDNA-based assay for benign and malignant gastrointestinal diseases [14] due to their discovery that high serum cfDNA levels were associated with other pathological conditions, such as rheumatoid arthritis [15] and pulmonary embolism [16].

Since then, a wide range of diagnostic approaches, collectively referred to as "liquid biopsies", has been developed [17,18]. While cancer-related approaches have often targeted circulating tumor DNA (ctDNA), carrying cancer-specific genetic alterations [19,20], other strategies have involved length profiling of cfDNA molecules [10] or screening for epigenetic modifications [21], which were specific for various malignancies [22]. Another well-established application for cfDNA-based liquid biopsies is the cell-free foetal DNA (cffDNA)-based prenatal test for the detection of Down syndrome and other trisomies in maternal blood [23–26].

Altogether, a large number of studies are currently focused on the diagnostic capabilities of cfDNA [17,18,22], while the characterization of the molecular mechanisms [27] underlying their release and biology remains largely neglected. Therefore, the purpose of this review is to outline the mechanisms of cfDNA release into the bloodstream and the biological properties of cfDNA molecules, as known thus far (Figure 1).



Cell-free DNA release mechanisms and types	
Apoptosis	• Nucleosomes
Necrosis Pyroptosis Autophagy Phagocytosis Mitotic catastrophe	• Mitochondrial DNA • Long cfDNA fragments
Erythroblast enucleation	• Erythroblast DNA
NETosis	• Neutrophil extracellular traps (NETs)
Macromolecular structures	• DNA-protein complexes • Virtosomes
Extracellular vesicles (EVs)	• Apoptotic bodies • Microvesicles • Exosomes
Various	• Single- and double stranded DNA
Micronucleation by chromosomal instability	• Extrachromosomal circular DNA • Other

Figure 1. Different forms of cell-free DNA (cfDNA) in the human circulatory system. The biological and structural characteristics of the aggregate cfDNA population in a typical biospecimen is highly heterogeneous. While some overlap is common (see the main text), different cellular sources and mechanisms of origin often result in the production of uniquely distinct forms of cfDNA. Exhaustive stratification of the different cfDNA subtypes and an improved understanding of the factors which might modulate these characteristics of cfDNA are vital steps toward understanding the biological role(s) of cfDNA, as well as an expansion of their clinical utility.

2. Apoptosis and Necrosis

For almost 20 years, apoptosis has been considered to be the primary source of circulating cfDNA in both healthy and diseased individuals [28–30], taking into account that about 50 to 70 billion cells perish via programmed cell death in the human body every day [31]. The maintenance of the homoeostatic balance and a steady turnover of cells are continuous tasks in multicellular organisms, as an imbalance between the formation of new cells and the removal of abnormal/blemished or superfluous cells can result in significant complications. Processes, such as oxidative stress and DNA damage, can trigger an imbalance in the cellular homeostasis, leading to the initiation of the intrinsic (mitochondrial) apoptotic pathway [32], which is controlled and carried out by the pro-apoptotic and pro-survival members of the BCL-2 family of proteins [33].

In contrast, the extrinsic apoptotic pathway is activated by the binding of extracellular signals ("death factors") to their respective receptors ("death receptor") [32] or, provided that specific conditions are met, to pattern recognition receptors (PRR) [34]. As different as the two pathways are, they both result in activation of a caspase-dependent proteolytic cascade, which is executed by initiator caspases (caspase-8 and -9) and, subsequently, executioner caspases (caspase-3, -6, and -7), which leads to inhibition of the inflammatory response [35], enzymatic degradation of cell components, DNA fragmentation, translocation of phosphatidylserine to the cell surface, formation of membrane blebs [36], and finally, packaging and release of extracellular vesicles (EVs) from apoptotic cells [37]. Inhibited or excessive apoptosis levels are connected to several medical conditions, including cancer, autoimmune, and neurodegenerative diseases [33].

The process of apoptotic nucleosomal DNA fragmentation starts with the concerted action of caspase-activated DNase (CAD), which is also termed DNA fragmentation factor subunit beta (DFFB), and DNaseI L-3 (or DNase γ), which translocates from the endoplasmatic reticulum to the nucleus when apoptosis has been initiated [38,39]. This DNase is involved in the degradation of intra- and extracellular DNA [40]. Since CAD is an endonuclease, lacking exonuclease activity, it cleaves the dsDNA of inter-nucleosomal linker regions [41,42], which is mirrored in nucleosomal footprints [43]. Extracellular DNaseI, which cleaves naked DNA that is not densely organized or protein-associated, is believed to be implicated in the "trimming" of cfDNA fragments to less than 150 bp in size [44] among other functions [40,45]. Watanabe and colleagues showed, in mouse experiments, that the deletion of CAD produced a single 180 bp DNA band, whereas the deletion of both CAD and DNaseI L-3 resulted in a decrease in cfDNA concentration and generation of cfDNA fragments of varying sizes (shorter or longer than mono-nucleosomal fragments and multiples thereof). A double deletion mutant, combining CAD and DNaseI L-3 deficiency, led to the complete absence of cfDNA [46]. Apart from DNaseI, further cfDNA degradation is executed partially in the bloodstream, for example, by plasma factor VII activating protease (FSAP) [47], which increases the accessibility of the DNA and Factor H [48], or by the liver, spleen, or kidney, as well as by affiliation with other blood constituents [49–52].

Apoptotic EVs are the main instrument to facilitate apoptotic cell clearance and to communicate with their environment, thereby initiating or inhibiting immune responses [37]. They comprise apoptotic bodies [53], apoptotic microvesicles, apoptotic exosomes [54], and apoptotic exosome-like particles [55]. Apoptotic bodies include proteins, lipids, (mi)RNA, and DNA [56–59]. Their production is believed to occur in a highly regimented manner [60,61]. According to Fernando et al. [62], more than 90% of cfDNA is related to exosomes, being located either on the surface of the vesicles or within the vesicle lumen [63].

Necrosis, i.e., accidental cell death, is a faster process than apoptosis which also contributes, in some instances, to the pool of cfDNA, mainly in the form of larger fragments (>1000 bp), but also short fragments, that are attributed to partial plasma nuclease cleavage of larger circulating fragments [64–67].

3. Erythroblast Enucleation

Roughly two million red blood cells (RBCs) with a half-life of approximately 115 to 120 days [68–70] mature every second in healthy adults [70,71]. Their complex biogenesis, starting from multipotent hematopoietic stem cells [71,72], involves several intermediate stages and occurs mostly in bone marrow, but also in the spleen and liver [71,73]. While nuclei of vertebrate RBCs become highly condensed, pyknotic, and transcriptionally inactive during maturation in general [69], mature mammalian erythrocytes even lack a nucleus [70–72]. This was discovered in 1875 by Gulliver [74] and was thought to enable higher hemoglobin levels in the blood [69,72]. The process of enucleation (formerly also called denucleation), which was visualized via electron microscopy by Simpson in 1967 [75], occurred at the stage of orthochromatic erythroblasts at the erythroblastic islands [70,71,73], which were discovered by Bessis, in 1958 [76]. Erythroblastic islands consist of a central macrophage, surrounded by physically attached erythroblasts, which mature from proerythroblasts via basophilic and polychromatophilic forms to orthochromatic erythroblasts, losing all organelles in the process [70,71,73]. Orthochromatic erythroblasts finally exit their cell cycle and enucleate, separating into a reticulocyte and into a pyrenocyte [70–72]. Reticulocytes contain most of the cytoplasm [77] and enter into the bloodstream, where they mature within a time frame of approximately two days [71]. Pyrenocytes consist of the nucleus which is surrounded by a thin rim of cytoplasm and the plasma membrane [78]. They are readily phagocytized by the central macrophage [70,71,73], which is promoted by phosphatidylserine residues on their surface, acting as "eat-me" signals [71,72,79]. Engulfed nuclei are subsequently digested by DNaseII in lysosomes [72,80].

Although, in the literature, erythroblast enucleation has been repeatedly suggested to be a potential source of cfDNA [27,81–83], experimental evidence for this hypothesis is relatively sparse. When analyzing blood plasma of sex-mismatched bone marrow transplantation patients via quantitative real-time PCR (qPCR), Lui et al. (2002) [84] found that plasma cfDNA was predominantly of donor origin, and they concluded that cfDNA stemmed largely from cells of the hematopoietic lineage. This conclusion was confirmed by methylation-based tissue-mapping approaches [85,86] and by nucleosome footprint analyses [43], where it was stated that approximately 55% of cfDNA originated from white blood cells and 30% from erythrocyte progenitors [86,87]. Exercise-induced elevation of cfDNA levels in the blood was also attributed to cells of the hematopoietic linage by Tug et al. (2015) [83] in an analysis of cfDNA from sex-mismatched hematopoietic stem cell recipients via qPCR. Using high-resolution methylation profiles, Lam et al. (2017) [88] identified three genomic loci featuring erythroblast-specific low methylation density. They determined that erythroid DNA represented about 30% of the plasma cfDNA via qPCR and postulated that degraded erythroblast DNA from bone marrow could somehow escape into circulation. The question of how exactly DNA from pyrenocytes relates to cfDNA fragments, which can be found in the blood plasma or serum, remains largely unanswered thus far.

4. NETosis

Another mechanism, the release of neutrophil extracellular traps (NETs), which can lead to the creation of cfDNA, was discovered more recently [89]. NETs have been characterized as an innate response of neutrophils, with the function to trap and kill microorganisms. According to Yipp et al. [90], the indications for this process, which have been termed NETosis [91,92], were observed earlier, but not named or recognized as an independent event. The NETosis process is now accepted as an independent reaction to a threat and is considered to be different from apoptosis and necrosis.

NETs consist of disintegrated chromatin, which "trap" the microorganisms, and also serves as an anionic binding matrix for different antimicrobial proteins in order to kill the entrapped bacteria [90,93,94]. The proteins are usually found in granules inside of the cell and are released during NETosis. This release can occur either slowly, via a lytic cell death, or rapidly, via an explosive discharge of disintegrated chromatin and peptides, while the cell remains functional and can still respond to the surroundings [93,94]. These two different forms of NETosis are termed suicidal and vital [90], respectively. The mechanism of the NET formation was initially considered to be solely the result of NADPH oxidase-dependent production of reactive oxygen species (ROS) [91], but evidence for NADPH independent pathways was discovered subsequently [95,96]. There was also evidence that cell cycle proteins could be regulating NET formation [97].

Persistent NET structures in the bloodstream can lead to vascular occlusion [40], therefore, an effective removal system is required. It has been shown that the degradation of the released chromatin starts by host DNases circulating in the body fluid. The degradation is performed by two independent DNases, namely, DNaseI and DNaseI L-3 [40]. This breakdown of the chromatin is not the only process of NET removal, as the full clearing is performed with the support of macrophages [98]. It seems that DNases are involved with preprocessing the NET structure, while the total NET removal is performed by macrophages and other cell types incorporating and digesting the remainders of the NETs.

There is an indication that NETs may have an influence on cfDNA levels, since large amounts of DNA are released into the bloodstream, or into tissues, in a relatively short time. This influence has been studied for different health conditions and can correspond to increased cfDNA levels in diseased patients [99]. The cfDNA levels are also influenced by stalling the degradation process. Several diseases have been shown to lead to problems with a regulated full degradation of NETs, thus leading to higher cfDNA levels [100–103].

In addition to the degradation by endogenous DNases, it has also been shown that different species of bacteria could escape NETs by degrading the chromatin matrix via the release of DNases [104–107]. It may be possible that bacterial DNases produce different fragments of NETs as compared with the endogenous DNase degradation system, which could lead to an altered degradation pattern. To date, there is a lack of evidence for all of these possibilities. To protect the NET structures and preserve their antimicrobial character, the immune system coats the chromatin fraction with peptides, such as LL-37 and defensin-3 against degradation, which then hinders the activity of these DNases [108].

NETs do not always have beneficial effects during an immune response and the response to a threat by NETosis can even have detrimental effects for the organism. It has been shown, for example, that dormant cancer cells can be reactivated during inflammation reactions involving NETs in mice [109]. In chronic inflammation reactions, the lytic cell death mechanism has been shown to be proinflammatory, and therefore could counteract medical treatments [110]. It was also shown by Katkar et al. [111] that an exaggerated immune response, containing NETs together with the lack of DNase activity, was the main cause of tissue destruction by snake venom in mice. Tissue destruction due to NETosis and the following immune response were shown in this particular response, and also in other instances [99]. The same response pattern, together with increased cfDNA levels, has also been shown in patients with active COVID-19 infection, as recently reported by Zuo et al. [112].

5. Macromolecular Structures

Gahan and Stroun [113] coined the term virtosome, in 2010, to describe a circulating DNA-RNA-lipoprotein complex, which was actively released from living cells. Active release of DNA was first observed for stimulated lymphocyte cells [114]. Subsequently, DNA release was also shown for non-stimulated rat and human lymphocytes [115,116]. Furthermore, DNA release was observed for different eukaryotic cell types, for example, frog heart auricles [117] or rat spleen [118] mostly using in vitro studies, but it was also observed in vivo, for example using chick embryo fibroblasts [119]. The release was dependent on the concentration of the complex in the medium, with high concentrations in the environment suppressing further release [115,120]. The release mechanism itself was mostly unknown, however it was suggested that it may appear in phase G0 or G1 of the cell cycle [113].

It has been confirmed previously, by means of a ^{3}H-thymidine labeling study, that DNA was newly synthesized within the cell [115]. The presence of proteins and lipids after treatment of the complex with proteinase K or lipase was observed and showed RNase activity [118]. A recent study by Cataldi and Viola-Magni [120] quantified the composition of virtosomes released from human lymphocytes into both the cytoplasm and the cell culture supernatant. This study found that the complexes in the cytoplasm contained approximately 3.45% DNA, 35.09% RNA, 19.90% phospholipids, and 41.01% proteins. Correspondingly, the cell culture supernatants contained 3.92% DNA, 36.41% RNA, 32.21% phospholipids, and 27.44% proteins. It was also suggested that there was no membrane around the virtosome [121], which was further supported by the low proportion of cholesterol and phosphatidylcholine in the complex [120]. These DNA molecules were believed to be around 450–700 bp long [113]. To our knowledge, there have not yet been any reports describing sequencing results (DNA and RNA) of virtosome complexes, which could be used to shed light onto the composition and specific origin of this fraction of cfDNA. Cells can uptake the virtosome complex [113] that has been released from cells of different cell types and the virtosomes can modify the biology of the receiving dividing cells [122,123]. This indicates that the virtosome complexes may be involved in signaling pathways between different cell types or horizontal gene transfer.

It is worth noting that Gahan and Stroun [113] pointed out several similarities between virtosomal DNA and metabolic DNA. It has been suggested that this unique population of DNA results from unscheduled DNA synthesis and occurs independently from the high molecular weight chromatin mass in cell nuclei [27,124]. In an early study of adult mouse heart, intestine, and skeletal muscle, it was found that metabolic DNA ranged between 500,000 and 4,000,000 Daltons [125] which equated to approximately 769 bp–6 Kbp in size. Interestingly, cfDNA fragments which exhibited this size profile have been observed in the supernatant of various cultured cell lines [126–128] and have been shown to be part of the cargo of some EVs [62,129,130]. They are often encountered in human biospecimens [131–133]. While it is an interesting possibility that metabolic DNA could serve as the precursor to larger ~1–6 Kbp cfDNA fragments, more research is needed to determine whether they truly represent a population of cfDNA originating from a regulated extrusion pathway, or whether they are simply the product of passive release mechanisms, such as apoptosis or necrosis.

6. Extracellular Vesicles

The release of EVs by cells was first described in 1967 for chondrocytes [134] and blood platelets [135]. These vesicles are characterized as single-lipid bilayer membrane extracellular organelles of simple spheroid morphology, which are present in all biological fluids tested thus far [136–138]. Constitutive release of EVs was reported for prokaryote [139] and eukaryote [134–138] cells and was considered to be a part of the normal physiology of all cells [138,140]. Currently, three main types can be distinguished, based on their size and their mechanisms of biogenesis [136,137,141] (Figure 2). Apoptotic bodies with diameters of 500–5000 nm are formed in the course of cell disintegration during programmed cell death [53,142]. Ectosomes, which encompass microparticles, microvesicles, and large vesicles with diameters from 50 to 1000 nm, are released directly from the plasma membrane via outward budding and fission [142,143]. Exosomes, with diameters of 30–200 nm, have been released via fusion of multivesicular bodies (MVBs) with the plasma membrane [136,137,141,143], which was first shown for rat reticulocytes, in 1983 [144]. The association of EVs with dsDNA was initially reported for human prostasomes, which were EVs secreted by the tissue of the prostatic gland, in a preliminary lab report from Olsson and Ronquist, in 1990 [145]. This result was finally confirmed almost twenty years later by Ronquist et al. (2009) [146]. In the following year, Guescini et al. (2010) [147], reported the presence of mitochondrial DNA (mtDNA) in EVs, which were characterized as exosomes released from astrocytes and glioblastoma cells. In prokaryotes, the release of DNA-containing outer membrane vesicles (OMVs) was first described, in 1989, by Dorward et al. [148] for *Neisseria gonorrhoeae* and by Garon et al. [149] for *Borrelia burgdorferi*. The studies also both showed that the encapsulated DNA was protected from digestion by DNaseI. When screening for viruses in hyperthermophilic archaea of the order Thermococcales, Soler et al. (2008) [150] observed the release of spherical membrane vesicles containing cellular DNA, which were not related to any viral activity and also exhibited resistance to DNase digestion.

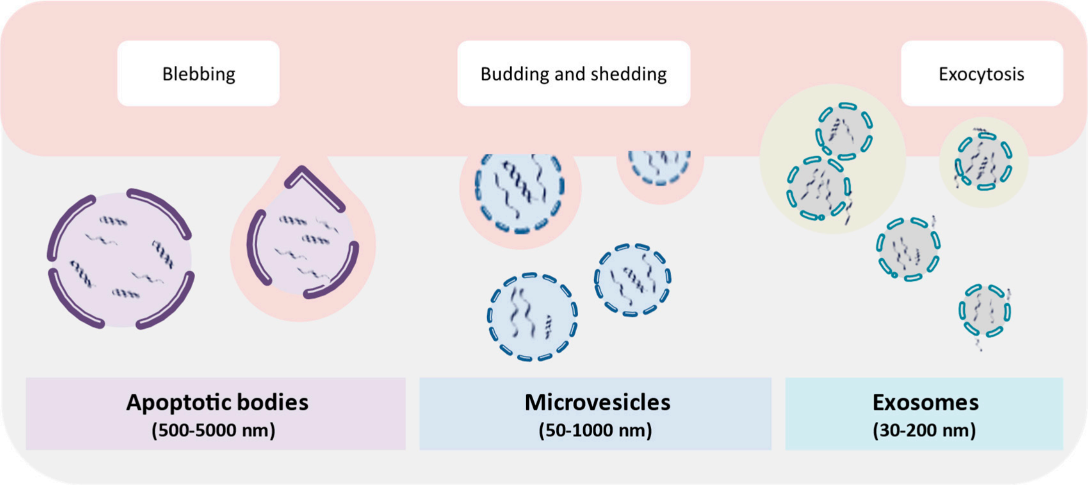


Figure 2. Different extracellular vesicle (EV) types in the human body. The three main types of EVs that occur in human body fluids include microvesicles, apoptotic bodies, and exosomes. These vesicles can be differentiated based on the mode of their production, cellular extrusion pathways, and their overall size. In addition, EVs can be stratified according to their content and function.

A number of studies have demonstrated that DNA was affiliated with exosomes [129,151–153] and microvesicles [154,155], which were, for example, derived from diverse tumor types [129,151,152,154,156] or normal cells [155–158]. Furthermore, it has been shown that DNA molecules originating from both healthy [155,157,158] and tumor [153,154,159] cells could be transferred to other cells via EVs and may affect gene expression [155,157–159]. Similar observations have been made for bacterial DNA shuttled by OMVs [148,160]. Apart from affecting the gene expression of recipient cells, stable genomic integration of DNA transferred by vesicles was also shown in a proof-of-principle experiment by Fischer et al. (2016) [161]. Therefore, it is clear that EVs constitute a distinct, and perhaps biologically active, subpopulation of cfDNA molecules. However, several aspects concerning EV-associated cfDNA remain poorly understood, mainly because the phenomenon has, thus far, been underinvestigated and conflicting results have been reported in the literature.

One reason is that a fraction of the total cfDNA population in the blood plasma or serum, which is associated with EVs has not yet been clearly defined. An early study suggested that, rather than floating freely, the majority of cfDNA in the blood of cancer patients was contained within exosomes [152]. In support of this, Fernando et al. (2017) observed that up to 90% of the cfDNA in human blood was exosome-associated [62]. In contrast, Helmig et al. (2015) found that only about 5% of the total cfDNA population in blood collected from healthy individuals before and after physical exercise was associated with EVs [82]. Taking into consideration the relatively short half-life of free-floating cfDNA [87,162], the discrepancy between the reported proportions of vesicle-associated cfDNA in the above studies may be explained by different sample handling procedures. For example, in the study by Helmig et al., 2015, [82], the blood was analyzed directly after the blood draw, while the blood samples used in both the Kahlert et al. (2014) [152] and Fernando et al. (2017) [62] studies were taken at an external facility with a larger temporal delay before the analysis.

A second aspect is the question that, prior to its release from cells, it is not clear what fraction of the genomic DNA is packaged within the interior of the EVs vs. molecules that are attached to their exterior surface. In line with this, it is not known how much cfDNA can and does attach to different EVs after entering the extracellular space. In the above study by Helmig et al. (2015) [82], DNase treatment suggested that only about a quarter of the EV-associated DNA was contained within the vesicles. This was consistent with the findings of Fischer et al. (2016) [161] and Lazaro-Ibañez et al. (2019) [163] who reported that cfDNA was mainly associated with the outer membrane surface of EVs, with a smaller amount being contained within the vesicles. However, additional research is needed to determine the exact proportions in baseline conditions, as well as changes upon various stimuli.

Third, relatively little is known about the biological and physiological characteristics of EV-associated DNA, including, but not limited to the following: (i) the mechanisms that underlie the vesicle-mediated release of DNA molecules, (ii) the relative proportion of the total EV-associated cfDNA population that is complexed with each of the different EV types (i.e., apoptotic bodies, microvesicles, and exosomes), (iii) the physiological factors that modulate these mechanisms, and ultimately (iv) the relative contribution of different cell types toward the total EV-associated cfDNA population. Despite the lack of research on these topics, some recent studies provide at least some insights. Takahashi et al. (2017) [164], described the sustained exocytosis of DNA fragments via exosomes as an important physiological process for the preservation of cellular homeostasis, where the inhibition of exosome secretion led to intracellular accumulation of cytosolic DNA. This resulted in the activation of the ROS-dependent DNA damage response, followed by senescence and ultimately cell death. They also observed an increased release of DNA via exosomes in senescent cells. This suggests that cytosolic DNA fragments can be a significant source of intracellular stress, which is connected to increased exosome secretion, as described by Hessvik and Llorente (2017) [136]. MVBs can either be directed to lysosomes, where their content is degraded, or to the cell membrane, where they fuse with the membrane and release their vesicular content into the extracellular space as exosomes [136,142]. The destination of MVBs is considered to be dependent on cellular homeostasis [136,165,166]. Such an increase in exosome secretion in connection to a stress response has also been described for bacteria [160]. It is also noteworthy that Brahmer et al. (2019) [167] reported an increased release of EVs triggered by exercise. They identified endothelial cells, platelets, as well as different types of leukocytes, to be the main sources of these vesicles. Since the DNA cargo of the vesicles was not addressed in the study, the question of whether the increase in vesicles could lead to an increased amount of vesicle-associated DNA in the bloodstream remains currently unanswered.

Fourth, and in line with the previous point, the composition and size of the cfDNA molecules that are associated with EVs is unclear. Studies have shown that EVs contain fragments of single-stranded DNA [154], double-stranded DNA [129,151,152], as well as mtDNA [147,153]. These fragments have been found to range from as small as ~100 bp up to sizes exceeding 10 Kbp [129,152,154]. While the excreted cytosolic DNA fragments have long been thought to be a product of DNA damage [159,164], the heterogeneous composition of the exosomal DNA content, which features disproportionally large amounts of DNA from retro-transposable elements (RTEs) and satellite repeat DNA [154,168], suggests that other mechanisms could contribute to the abundance of cytosolic DNA fragments in a significant way. This overrepresentation of DNA from RTEs and satellite repeats has been found for cancer-cell culture supernatants [168,169], as well as the serum of healthy human subjects [169]. The respective fractions of these elements were significantly increased in human sepsis patients (bacterial and fungal sepsis), indicating an increased release in response to their medical condition [169].

7. Chromosomal Instability and Micronucleation

Recent reports have indicated the possibility that a portion of the total cfDNA population in vitro, as well as in vivo, could be a product of the effects of chromosome instability (CIN). CIN is one form of genetic instability that is characterized by sustained and often accelerated changes in the chromosome structure or number [170]. There are different pathways through which CIN results in the release of genomic DNA into the extracellular space, each of which can be modulated by a wide range of biological factors, thus, representing multiple layers of regulation [168]. Here, we limit the focus to specific chromosome mis-segregation events that can occur during mitosis.

7.1. Fundamental Chromosome Segregation Errors During Mitosis

After DNA replication, chromosomes possess two kinetochores, which are complex protein structures that serve as attachment points for the mitotic spindle. In the prometaphase, these kinetochores interact with spindle microtubules in both a lateral and end-on fashion. During the metaphase, each kinetochore binds to microtubules oriented towards opposite spindle poles. These bi-oriented kinetochore-microtubule (k-MT) attachments are crucial for the proper alignment and segregation of the chromosomes that are attached to the spindle [171]. However, due to the asynchronous and stochastic nature of the initial capture of microtubules by kinetochores, some k-MT attachments are not bi-oriented, resulting in either delayed or inappropriate attachment of chromosomes to the spindle during the prometaphase-to-metaphase transition [172]. Among the various k-MT attachment errors that can occur, merotelic attachments constitute a pathway for the generation of cfDNA. In merotelic attachments, one kinetochore binds to microtubules growing from opposite spindle poles [173,174]. Since merotelic attachments are not detected by the mitotic spindle checkpoint, cells can proceed to anaphase without completing error correction [175]. Although most merotelic chromosomes segregate correctly during anaphase, a small fraction remain at the spindle equator, resulting in lagging chromosomes [173]. Interestingly, if there is sufficient distance between a lagging chromosome and the main chromatin mass at the end of cell division, the lagging chromosome can recruit its own nuclear envelope and form a so-called micronucleus (MN), which is also known as a Howell–Jolly body [176]. In such cases, an interphase daughter cell contains the following two types of nuclei: (a) the primary large nucleus and (b) up to several smaller micronuclei (MNi) that house the mis-segregated chromosomes. While chromosome lagging and micronucleation can occur at low levels in normal cells [175,177], increased levels of MNi have been shown for various pathologies, especially in cancer cells with CIN [178–182]. In keeping with this, we hypothesize that MNi, with their DNA cargo, may translocate to the extracellular space and serve as one of the sources of cfDNA.

7.2. Other Chromosome Mis-Segregation Events that Can Arise During Mitosis

Apart from fundamental chromosome segregation errors, such as merotely, MNi can also arise during mitosis as a result of telomere loss. Telomeres preserve genomic stability by protecting natural chromosome ends from degradation, illegitimate self-recombination, and end joining with nearby chromosomes [183]. However, in certain cases (e.g., cancer), accelerated shortening of telomeres (loss of the terminal sequence), or complete loss of the end-capping structure is a prevalent feature [168,184–187]. Loss of telomeres is mainly the consequence of the following: (i) attrition of telomere repeats, which is typically associated with repression or reduced activity of telomerase [188]; (ii) loss of specific telomeric proteins, which may prompt the cell to identify chromosome ends as DNA breaks [189]; or (iii) double-stranded DNA breaks (DSBs), which may form as a result of hypomethylation followed by transposon-induced breaks [190–192], mis-repaired breaks caused by radiation [193–196], and deprivation of important metabolites, for example, folate [197,198]. When a telomere-deficient chromosome is replicated, the two ends of the sister chromatids fuse and form a chromosome with two centromeres, i.e., a dicentric chromosome [199,200]. Since a dicentric chromosome is able to attach to both spindle poles, the two centromeres are pulled to opposite poles during the anaphase, forming a continuous string of chromatin, stretching from one pole to the other, normally referred to as an anaphase bridge [201]. Anaphase bridges often break, resulting in various chromosomal abnormalities, often followed by the formation of MNi. The characteristics of the MNi and its DNA content depend on the location of the breaking point in the anaphase bridge [202–205].

Repeated pulling apart of dicentric chromosomes, anaphase bridge formation, and subsequent breakage in gene regions over multiple cell divisions, also known as breakage-fusion-bridge (BFB) cycles, results in the amplification of DNA sequences that are adjacent to the break or fusion point [201,206]. The BFB cycles are typically sustained until the chromosome acquires a new telomere. However, during the process, recombination of homotypic sequences within the amplified DNA often results in the formation of mini circles of acentric and atelomeric DNA, which are eliminated from the aberrant chromosome. These structures, which are known as double minutes (DMs) [201,206], are capable of replication and can also localize to the nuclear periphery, exit the nucleus through budding, and eventually become extruded from cells in the form of microcells [203,207]. When they are in the form of nuclear buds (NBUDs) after exiting the nucleus, they have a similar morphology as the typical MN [202]. Interestingly, this extra-chromosomally amplified DNA frequently consists of oncogenes [208–210]. This process, which is summarized in Figure 3, may explain the recently reported presence of extrachromosomal circular DNA in the circulatory system of both mice and humans [211–213]. It may also explain the significant overrepresentation of specific retrotransposons in the cfDNA isolated from the cell culture supernatant of human bone osteosarcoma (143B) cells [168].

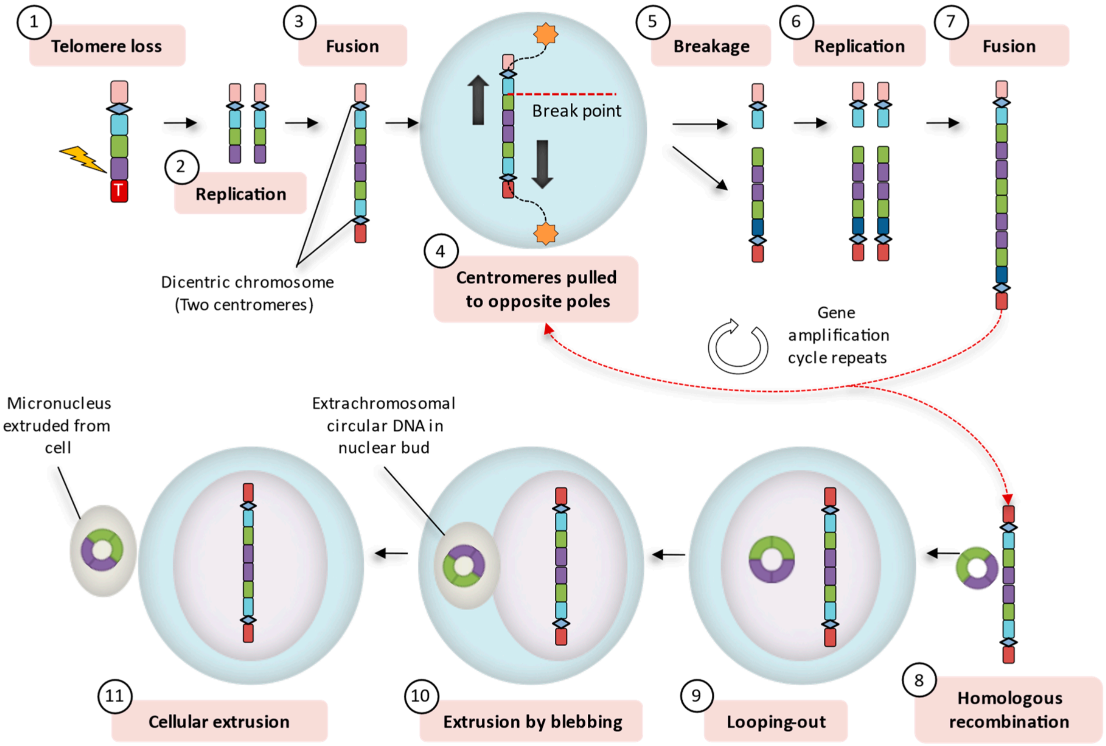


Figure 3. Breakage-fusion-bridge cycles as a potential source of cell-free DNA (cfDNA). A chromosome that has lost its telomere (**1**) is replicated (**2**) and the sister chromatids fuse at the broken ends, forming a dicentric chromosome (**3**). The two centromeres are, then, pulled to opposite spindle poles during the anaphase, stretching out the chromatin to form an anaphase bridge (**4**). In many cases an anaphase bridge breaks (**5**) and the broken telomere-deficient chromosomes are again replicated (**6**) and once again fuse with their sister chromatids (**7**). Several repetitions of this process result in the amplification of DNA sequences that are adjacent to the break or fusion point. An anaphase bridge may break in various regions, resulting in the amplification of different genomic regions (e.g., genes, oncogenes, and repetitive DNA). Recombination between homotypic sequences within the amplified DNA generates extrachromosomal circular DNA (**8**), which can eventually become looped out of the chromosome (**9**). These mini circles of DNA can subsequently be extruded from the nucleus through budding (**10**) and eventually become extruded from cells into the extracellular space (**11**).

Anaphase bridges can also break in non-gene regions. Differentiation between chromosomes with varying telomere erosion profiles showed that the long and short arms of chromosomes 1 and 22, respectively, had critically short telomeres as compared with other chromosomes. Consistent with this observation, telomeric fusion of these chromosomes was the most striking cytogenetic abnormality observed at anaphase after the first population doubling of human mammary epithelial cells. However, since chromosome 1 contributed more to the formation of anaphase bridges, it was suggested that the loss of telomere function did not modulate the formation of nuclear anomalies through a single mechanism [214]. This concurred with a previous study showing that telomere dysfunction caused a spectrum of mitotic defects [194]. In this study, two different types of anaphase bridges were observed, i.e., one type consisting of one or two strings of decondensed chromatin connected to both spindle poles, whereas the other bridge had detached from either one or both spindle poles (resulting in lagging of entire chromosomes). Fragmented DNA was localized either in NBUDs, which contained completely fragmented and broken anaphase bridges, or in internuclear strings, which originated from anaphase bridges that were fragmented, but not fully broken. Interestingly, in more than 90% of strings and protrusions, DNA fragmentation began adjacent to the centromere and continued distally. Karyotype analysis of 338 cases of colorectal adenocarcinomas confirmed that this pericentromeric breakpoint pattern also occurred in vivo [194].

In addition to the above observations, molecular studies of immunodeficiency, centromeric region instability, and facial anomalies (ICF syndrome) have shown that chromosomes 1 and 16 and, in some cases, chromosomes 2 and 9 were predisposed to hypomethylation at their centromeric and pericentromeric regions [215], resulting in a variety of cytogenetic abnormalities, which included the extrusion of self-associated satellite DNA into MNi or NBUDs [216]. In line with this, several cancer types have shown preferential hypomethylation of the satellite repeat arrays on the basis of the q-arms of chromosomes 1, 9, and 16 [217–219]. This hypomethylation was causally linked to a high frequency of non-random rearrangements in the centromeric and pericentromeric heterochromatin of these chromosomes, which were nearly identical to those observed in ICF syndrome. In support of this, similar genomic aberrations and MNi-containing DNA derived from the centromeric and pericentromeric regions of chromosomes 1, 9, and 16 were induced in cultured cells through exposure to various clastogenic compounds that prevent normal methylation [220–224]. Taken together, these studies may explain why a growing number of studies have reported an overrepresentation of repetitive DNA in cfDNA [168,169,225–227]. More specifically, these studies may provide an explanation for the question why cfDNA derived from cultured cancer cells has been found to be enriched in repetitive DNA, which originated from the centromeric and pericentromeric regions of chromosomes 1, 9, and 22 [168].

While these studies provided indirect evidence that a portion of the cfDNA originate from CIN-induced MNi, there is currently no direct experimental evidence that shows this. Indeed, as far as we know, there have been, thus far, no attempts to isolate "free-floating" MNi directly from human body fluids. Therefore, further studies on the nature of MNi would likely provide deeper insight into the biology and physicochemical properties of cfDNA.

8. Discussion

Cell death, primarily via apoptosis or, under certain circumstances, necrosis, has often been considered to be the only relevant mechanism of DNA release into the bloodstream. This assumption has been justified by the non-random fragmentation pattern of circulating cfDNA [44,66,67]. In contrast, several active mechanisms for the release of cfDNA from cells have now also been described, where, for instance, a considerable fraction of the total cfDNA, which is attributed to the erythroid cell lineage [85–87,228], cannot sufficiently be explained by cell death, since mature erythrocytes do not possess a nucleus when undergoing eryptosis (erythrocyte apoptosis) [229]. Additionally, the mechanism of NETosis, which contributes substantially to the cfDNA fraction is not considered to be a passive mechanism [92,230]. Active release of DNA has been observed in many parts of the eukaryotic kingdom, i.e., it has been observed, for example, in humans, rats [231], chickens [232] and *allomyces arbusculus* (fungi) [233]. However, in these papers, there were no detailed characterizations described which dealt with the mechanism of cfDNA or the associated functions. This led Elzanowska et al. (in 2020) [234] to conclude that "DNA can also be released from living cells by active cellular secretion, although at present little is known about why functional cells secrete DNA and what is the biological significance of this process.

An entirely random release of DNA into the bloodstream as the predominant mechanism of cfDNA generation (i.e., essentially resulting fragmented entire chromosomes entering the bloodstream) contradicts, in our opinion, the fact that exosomal cfDNA clearly has an effect on cultured cells. For example, when exosomes hailing from radiated cell cultures were inoculated into cell cultures, which were not radiated, this led to a similar phenotype in the non-radiated cells as in the radiated cells [235–238]. In addition, the fact that exosomal cfDNA can be incorporated into the nuclei of inoculated cells [161] within a short time frame allows us to speculate that the cfDNA molecules contained in these exosomes may contain messages, which can be utilized in the uptaking cell.

Passively released cfDNA is still a useful tool, which can be used in a biomedical context for certain diseases, although, in our opinion, this is limited to diseases which are connected to dramatic changes to the genome of the patient (e.g., chromosomal rearrangements in tumors). Currently, only a small number of diagnostic tests using cfDNA are approved by the U.S. Food and Drug Administration, for example, the cobas® EGFR Mutation PCR Test v2 for non-small cell lung cancer [239], a test for ctDNA for PIK3CA mutated hormone receptor positive breast cancer [240], and the SEPT9 methylated DNA test for colorectal cancer [241]. The most prominent example is certainly the prenatal diagnostic test for Down syndrome and other chromosomal abnormalities [23–26], which is, to our knowledge, currently the only widely used cfDNA based diagnostic procedure available in public health institutions around the globe. Several groups are also developing screening tools for resistance in cancers, using cfDNA fragments originating from tumor tissue and constructing maps of tumor chromosomes after high-throughput DNA sequencing to determine mutations (insertions/deletions) on the cancer chromosomes [242–244]. As the detection limits for this cfDNA fraction are lowered (currently they lie around 2–5% of the total cfDNA), this is becoming a very valuable tool for non-invasive diagnosis of cancers [66,245].

Elevated levels of circulating cfDNA have been reported for several forms of physiological stress such as different disease conditions [12,13,15,99,246] or even simply exhaustion from physical exercise [82,83,247]. Correspondingly, increased active release of DNA fragments from cells via exosomes was found in connection with intracellular stress and senescence [164]. It is, in our opinion, unlikely that these cfDNA fractions are solely the product of apoptosis and necrosis, as very often the increase in cfDNA levels in the bloodstream occurs shortly after the stress exposure and is quite substantial [162]. The interplay of multiple cfDNA release mechanisms makes it difficult to relate the profile of cfDNA in blood to distinct diseases. We have experienced this ourselves when we investigated the occurrence of certain cfDNA sequence motifs (biomarkers) in relation to postsurgical bacteremia and sepsis [248]. Evaluating the frequencies of the motifs relative to the total cfDNA concentration of the samples did not yield any reliable correlation to the disease state. Therefore, we had to assume that several mechanisms contributed to the total cfDNA population being also overlaid (and thus partially masked) by the response to non-disease-related physiological causes, thus, interfering with our attempts to normalize the disease signals.

Evaluating the motif frequencies relative to each other in the form of motif pairs provided an intrinsic reference system, which ultimately enabled a motif-based distinction between samples from sepsis patients and samples from healthy probands [248]. These results provided evidence that information may be contained in cfDNA fragment frequency, methylation patterns, or the occurrence of mutations in some genomic regions, and also in the relative abundance of particular cfDNA motifs. The cfDNA fraction comprised by certain sequence elements, especially RTE and satellite repeats, seemed to change in response to physiological conditions [169,248], which could provide novel perspectives for future diagnostic approaches. Similar overrepresentation of these repeat elements has also been found in human bone osteosarcoma cancer cell culture supernatants [168]. Since different methods for DNA purification, amplification, and sequencing were used in these studies, it is very unlikely that these results were obtained due to artifacts caused by the applied methodology. Assessing the contribution of different cfDNA release mechanisms under varying physiological conditions may be crucial for the identification of additional reliable and robust signals within the information continuum provided by the cfDNA composition in body fluids. We expect that this should lead to the development of further diagnostic assays in the future. This research and development field is only emerging now, with very detailed studies possible due to third-generation high-throughput DNA sequencing methods being applied to the characterization of the entire DNA content of serum or plasma samples from patients and controls. Unlike the diagnostic assays, which solely consider passively released DNA as the target, the efforts focusing on including actively released cfDNA molecules have the advantage that they deal with the full spectrum of DNA circulating in the bloodstream, thus, increasing the signal-to-noise ratio considerably.

Author Contributions: Conceptualization, S.G., C.W.S., and A.J.B.; writing—original draft preparation, S.G., A.J.B., S.H., I.R.R., K.P.S., D.S., V.U., and C.W.S.; writing—review and editing, S.G., A.J.B., S.H., I.R.R., K.P.S., D.S., V.U., and C.W.S.; Figures, A.J.B. and V.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding, except for Open Access Funding by Graz University of Technology.

Acknowledgments: A.J.B. is supported by the Georg Forster Research Fellowship Program of the Alexander von Humboldt Foundation. The financial assistance of the Alexander von Humboldt Foundation is hereby acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mandel, P.; Metais, P. Les acides nucléiques du plasma sanguin chez l'homme. *CR Seances Soc. Biol. Fil.* **1948**, *142*, 241–243.
2. Hui, L.; Maron, J.L.; Gahan, P.B. Other body fluids as non-invasive sources of cell-free DNA/RNA. *Adv. Predict. Prev. Pers. Med.* **2015**, *5*, 295–323. [CrossRef]
3. Zachariah, R.R.; Schmid, S.; Buerki, N.; Radpour, R.; Holzgreve, W.; Zhong, X. Levels of circulating cell-free nuclear and mitochondrial dna in benign and malignant ovarian tumors. *Obstet. Gynecol.* **2008**, *112*, 843–850. [CrossRef] [PubMed]
4. Kohler, C.; Radpour, R.; Barekati, Z.; Asadollahi, R.; Bitzer, J.; Wight, E.; Bürki, N.; Diesch, C.; Holzgreve, W.; Zhong, X.Y. Levels of plasma circulating cell free nuclear and mitochondrial DNA as potential biomarkers for breast tumors. *Mol. Cancer* **2009**, *8*. [CrossRef]
5. Long, Y.; Zhang, Y.; Gong, Y.; Sun, R.; Su, L.; Lin, X.; Shen, A.; Zhou, J.; Caiji, Z.; Wang, X.; et al. Diagnosis of Sepsis with Cell-free DNA by Next-Generation Sequencing Technology in ICU Patients. *Arch. Med. Res.* **2016**, *47*, 365–371. [CrossRef]
6. Kowarsky, M.; Camunas-Soler, J.; Kertesz, M.; De Vlaminck, I.; Koh, W.; Pan, W.; Martin, L.; Neff, N.F.; Okamoto, J.; Wong, R.J.; et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 9623–9628. [CrossRef]
7. Stroun, M.; Anker, P.; Lyautey, J.; Lederrey, C.; Maurice, P.A. Isolation and characterization of DNA from the plasma of cancer patients. *Eur. J. Cancer Clin. Oncol.* **1987**, *23*, 707–712. [CrossRef]
8. McCoubrey-Hoyer, A.; Okarma, T.B.; Holman, H.R. Partial purification and characterization of plasma DNA and its relation to disease activity in systemic lupus erythematosus. *Am. J. Med.* **1984**, *77*, 23–34. [CrossRef]
9. Rumore, P.M.; Steinman, C.R. Endogenous circulating DNA in systemic lupus erythematosus. Occurrence as multimeric complexes bound to histone. *J. Clin. Investig.* **1990**, *86*, 69–74. [CrossRef]
10. Giacona, M.B.; Ruben, G.C.; Iczkowski, K.A.; Roos, T.B.; Porter, D.M.; Sorenson, G.D. Cell-free DNA in human blood plasma: Length measurements in patients with pancreatic cancer and healthy controls. *Pancreas* **1998**, *17*, 89–97. [CrossRef]
11. Thierry, A.R.; El Messaoudi, S.; Gahan, P.B.; Anker, P.; Stroun, M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev.* **2016**, *35*, 347–376. [CrossRef] [PubMed]
12. Leon, S.A.; Shapiro, B.; Sklaroff, D.M.; Yaros, M.J. Free DNA in the Serum of Cancer Patients and the Effect of Therapy. *Cancer Res.* **1977**, *37*, 646–650. [PubMed]
13. Tan, E.M.; Schur, P.H.; Carr, R.I.; Kunkel, H.G. Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. *J. Clin. Investig.* **1966**, *45*, 1732–1740. [CrossRef]
14. Shapiro, B.; Chakrabarty, M.; Cohn, E.M.; Leon, S.A. Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. *Cancer* **1983**, *51*, 2116–2120. [CrossRef]
15. Leon, S.A.; Ehrlich, G.E.; Shapiro, B.; Labbate, V.A. Free DNA in the serum of rheumatoid arthritis patients. *J. Rheumatol.* **1977**, *4*, 139–143. [PubMed]

16. Davis, G.L.; Davis Iv, J.S. Detection of circulating dna by counterimmunoelectrophoresis (cie). *Arthritis Rheum.* **1973**, *16*, 52–58. [CrossRef]
17. Jiang, P.; Lo, Y.M.D. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. *Trends Genet.* **2016**, *32*, 360–371. [CrossRef]
18. Wan, J.C.M.; Massie, C.; Garcia-Corbacho, J.; Mouliere, F.; Brenton, J.D.; Caldas, C.; Pacey, S.; Baird, R.; Rosenfeld, N. Liquid biopsies come of age: Towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **2017**, *17*, 223–238. [CrossRef]
19. Chan, K.C.A.; Jiang, P.; Zheng, Y.W.L.; Liao, G.J.W.; Sun, H.; Wong, J.; Siu, S.S.N.; Chan, W.C.; Chan, S.L.; Chan, A.T.C.; et al. Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin. Chem.* **2013**, *59*, 211–224. [CrossRef]
20. Chu, D.; Park, B.H. Liquid biopsy: Unlocking the potentials of cell-free DNA. *Virchows Arch.* **2017**, *471*, 147–154. [CrossRef]
21. Chan, K.C.A.; Jiang, P.; Chan, C.W.M.; Sun, K.; Wong, J.; Hui, E.P.; Chan, S.L.; Chan, W.C.; Hui, D.S.C.; Ng, S.S.M.; et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18761–18768. [CrossRef]
22. Suraj, S.; Dhar, C.; Srivastava, S. Circulating nucleic acids: An analysis of their occurrence in malignancies. *Biomed. Rep.* **2016**, *6*, 8–14. [CrossRef]
23. Dennis Lo, Y.M.; Corbetta, N.; Chamberlain, P.F.; Rai, V.; Sargent, I.L.; Redman, C.W.G.; Wainscoat, J.S. Presence of fetal DNA in maternal plasma and serum. *Lancet* **1997**, *350*, 485–487. [CrossRef]
24. Palomaki, G.E.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E.; Neveux, L.M.; Ehrich, M.; Van Den Boom, D.; Bombard, A.T.; Deciu, C.; Grody, W.W.; et al. DNA sequencing of maternal plasma to detect Down syndrome: An international clinical validation study. *Genet. Med.* **2011**, *13*, 913–920. [CrossRef]
25. Palomaki, G.E.; Deciu, C.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E.; Neveux, L.M.; Ehrich, M.; Van Den Boom, D.; Bombard, A.T.; Grody, W.W.; et al. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: An international collaborative study. *Genet. Med.* **2012**, *14*, 296–305. [CrossRef] [PubMed]
26. Pös, O.; Biró, O.; Szemes, T.; Nagy, B. Circulating cell-free nucleic acids: Characteristics and applications. *Eur. J. Hum. Genet.* **2018**, *26*, 937–945. [CrossRef] [PubMed]
27. Aucamp, J.; Bronkhorst, A.J.; Badenhorst, C.P.S.; Pretorius, P.J. The diverse origins of circulating cell-free DNA in the human body: A critical re-evaluation of the literature. *Biol. Rev. Camb. Philos. Soc.* **2018**, *93*, 1649–1683. [CrossRef] [PubMed]
28. Jahr, S.; Hentze, H.; Englisch, S.; Hardt, D.; Fackelmayer, F.O.; Hesch, R.D.; Knippers, R. DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **2001**, *61*, 1659–1665.
29. Sai, S.; Ichikawa, D.; Tomita, H.; Ikoma, D.; Tani, N.; Ikoma, H.; Kikuchi, S.; Fujiwara, H.; Ueda, Y.; Otsuji, E. Quantification of plasma cell-free DNA in patients with gastric cancer. *Anticancer Res.* **2007**, *27*, 2747–2751. [PubMed]
30. Delgado, P.O.; Alves, B.C.A.; Gehrke, F.d.S.; Kuniyoshi, R.K.; Wroclavski, M.L.; Del Giglio, A.; Fonseca, F.L.A. Characterization of cell-free circulating DNA in plasma in patients with prostate cancer. *Tumour Biol. J. Int. Soc. Oncodev. Biol. Med.* **2013**, *34*, 983–986. [CrossRef]
31. Reed, J.C. Dysregulation of apoptosis in cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **1999**, *17*, 2941–2953. [CrossRef] [PubMed]
32. Galluzzi, L.; Vitale, I.; Aaronson, S.A.; Abrams, J.M.; Adam, D.; Agostinis, P.; Alnemri, E.S.; Altucci, L.; Amelio, I.; Andrews, D.W.; et al. Molecular mechanisms of cell death: Recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ.* **2018**, *25*, 486–541. [CrossRef]
33. Singh, R.; Letai, A.; Sarosiek, K. Regulation of apoptosis in health and disease: The balancing act of BCL-2 family proteins. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 175–193. [CrossRef] [PubMed]
34. Amarante-Mendes, G.P.; Adjemian, S.; Branco, L.M.; Zanetti, L.C.; Weinlich, R.; Bortoluci, K.R. Pattern Recognition Receptors and the Host Cell Death Molecular Machinery. *Front. Immunol.* **2018**, *9*, 2379. [CrossRef]

35. White, M.J.; McArthur, K.; Metcalf, D.; Lane, R.M.; Cambier, J.C.; Herold, M.J.; van Delft, M.F.; Bedoui, S.; Lessene, G.; Ritchie, M.E.; et al. Apoptotic caspases suppress mtDNA-induced STING-mediated type I IFN production. *Cell* **2014**, *159*, 1549–1562. [CrossRef] [PubMed]
36. Martin, S.J.; Henry, C.M.; Cullen, S.P. A perspective on mammalian caspases as positive and negative regulators of inflammation. *Mol. Cell* **2012**, *46*, 387–397. [CrossRef]
37. Caruso, S.; Poon, I.K.H. Apoptotic cell-derived extracellular vesicles: More than just debris. *Front. Immunol.* **2018**, *9*. [CrossRef]
38. Errami, Y.; Naura, A.S.; Kim, H.; Ju, J.; Suzuki, Y.; El-Bahrawy, A.H.; Ghonim, M.A.; Hemeida, R.A.; Mansy, M.S.; Zhang, J.; et al. Apoptotic DNA fragmentation may be a cooperative activity between caspase-activated deoxyribonuclease and the poly(ADP-ribose) polymerase-regulated DNAS1L3, an endoplasmic reticulum-localized endonuclease that translocates to the nucleus during apoptosis. *J. Biol. Chem.* **2013**, *288*, 3460–3468. [CrossRef]
39. Koyama, R.; Arai, T.; Kijima, M.; Sato, S.; Miura, S.; Yuasa, M.; Kitamura, D.; Mizuta, R. DNase γ, DNase I and caspase-activated DNase cooperate to degrade dead cells. *Genes Cells* **2016**, *21*, 1150–1163. [CrossRef]
40. Jiménez-Alcázar, M.; Rangaswamy, C.; Panda, R.; Bitterling, J.; Simsek, Y.J.; Long, A.T.; Bilyy, R.; Krenn, V.; Renné, C.; Renné, T.; et al. Host DNases prevent vascular occlusion by neutrophil extracellular traps. *Science* **2017**, *358*, 1202–1206. [CrossRef]
41. Kitazumi, I.; Tsukahara, M. Regulation of DNA fragmentation: The role of caspases and phosphorylation. *FEBS J.* **2011**, *278*, 427–441. [CrossRef] [PubMed]
42. Enari, M.; Sakahira, H.; Yokoyama, H.; Okawa, K.; Iwamatsu, A.; Nagata, S. A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature* **1998**, *391*, 43–50. [CrossRef]
43. Snyder, M.W.; Kircher, M.; Hill, A.J.; Daza, R.M.; Shendure, J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **2016**, *164*, 57–68. [CrossRef] [PubMed]
44. Han, D.S.C.; Ni, M.; Chan, R.W.Y.; Chan, V.W.H.; Lui, K.O.; Chiu, R.W.K.; Lo, Y.M.D. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **2020**, *106*, 202–214. [CrossRef]
45. Jiménez-Alcázar, M.; Napirei, M.; Panda, R.; Köhler, E.C.; Kremer Hovinga, J.A.; Mannherz, H.G.; Peine, S.; Renné, T.; Lämmle, B.; Fuchs, T.A. Impaired DNase1-mediated degradation of neutrophil extracellular traps is associated with acute thrombotic microangiopathies. *J. Thromb. Haemost.* **2015**, *13*, 732–742. [CrossRef]
46. Watanabe, T.; Takada, S.; Mizuta, R. Cell-free DNA in blood circulation is generated by DNase1L3 and caspase-activated DNase. *Biochem. Biophys. Res. Commun.* **2019**, *516*, 790–795. [CrossRef]
47. Stephan, F.; Marsman, G.; Bakker, L.M.; Bulder, I.; Stavenuiter, F.; Aarden, L.A.; Zeerleder, S. Cooperation of factor VII-activating protease and serum DNase I in the release of nucleosomes from necrotic cells. *Arthritis Rheumatol.* **2014**, *66*, 686–693. [CrossRef]
48. Martin, M.; Leffler, J.; Smolag, K.I.; Mytych, J.; Björk, A.; Chaves, L.D.; Alexander, J.J.; Quigg, R.J.; Blom, A.M. Factor H uptake regulates intracellular C3 activation during apoptosis and decreases the inflammatory potential of nucleosomes. *Cell Death Differ.* **2016**, *23*, 903–911. [CrossRef]
49. Butler, T.M.; Spellman, P.T.; Gray, J. Circulating-tumor DNA as an early detection and diagnostic tool. *Curr. Opin. Genet. Dev.* **2017**, *42*, 14–21. [CrossRef]
50. Yu, S.C.Y.; Lee, S.W.Y.; Jiang, P.; Leung, T.Y.; Chan, K.C.A.; Chiu, R.W.K.; Lo, Y.M.D. High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. *Clin. Chem.* **2013**, *59*, 1228–1237. [CrossRef]
51. Celec, P.; Vlková, B.; Lauková, L.; Bábíčková, J.; Boor, P. Cell-free DNA: The role in pathophysiology and as a biomarker in kidney diseases. *Expert Rev. Mol. Med.* **2018**, *20*, e1. [CrossRef] [PubMed]
52. Bronkhorst, A.J.; Ungerer, V.; Holdenrieder, S. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol. Detect. Quantif.* **2019**, *17*, 100087. [CrossRef]
53. Kerr, J.F.; Wyllie, A.H.; Currie, A.R. Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* **1972**, *26*, 239–257. [CrossRef] [PubMed]
54. Renò, F.; Burattini, S.; Rossi, S.; Luchetti, F.; Columbaro, M.; Santi, S.; Papa, S.; Falcieri, E. Phospholipid rearrangement of apoptotic membrane does not depend on nuclear activity. *Histochem. Cell Biol.* **1998**, *110*, 467–476. [CrossRef] [PubMed]

55. Park, S.J.; Kim, J.M.; Kim, J.; Hur, J.; Park, S.; Kim, K.; Shin, H.J.; Chwae, Y.J. Molecular mechanisms of biogenesis of apoptotic exosome-like vesicles and their roles as damage-associated molecular patterns. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E11721–E11730. [CrossRef] [PubMed]
56. Bayraktar, R.; Van Roosbroeck, K.; Calin, G.A. Cell-to-cell communication: microRNAs as hormones. *Mol. Oncol.* **2017**, *11*, 1673–1686. [CrossRef]
57. Blander, J.M. The many ways tissue phagocytes respond to dying cells. *Immunol. Rev.* **2017**, *277*, 158–173. [CrossRef]
58. Crescitelli, R.; Lässer, C.; Szabó, T.G.; Kittel, A.; Eldh, M.; Dianzani, I.; Buzás, E.I.; Lötvall, J. Distinct RNA profiles in subpopulations of extracellular vesicles: Apoptotic bodies, microvesicles and exosomes. *J. Extracell. Vesicles* **2013**, *2*. [CrossRef]
59. Savill, J.; Fadok, V. Corpse clearance defines the meaning of cell death. *Nature* **2000**, *407*, 784–788. [CrossRef]
60. Atkin-Smith, G.K.; Tixeira, R.; Paone, S.; Mathivanan, S.; Collins, C.; Liem, M.; Goodall, K.J.; Ravichandran, K.S.; Hulett, M.D.; Poon, I.K.H. A novel mechanism of generating extracellular vesicles during apoptosis via a beads-on-a-string membrane structure. *Nat. Commun.* **2015**, *6*, 1–10. [CrossRef]
61. Atkin-Smith, G.K.; Poon, I.K.H. Disassembly of the Dying: Mechanisms and Functions. *Trends Cell Biol.* **2017**, *27*, 151–162. [CrossRef] [PubMed]
62. Fernando, M.R.; Jiang, C.; Krzyzanowski, G.D.; Ryan, W.L. New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes. *PLoS ONE* **2017**, *12*, e0183915. [CrossRef] [PubMed]
63. Torralba, D.; Baixauli, F.; Villarroya-Beltri, C.; Fernández-Delgado, I.; Latorre-Pellicer, A.; Acín-Pérez, R.; Martín-Cófreces, N.B.; Jaso-Tamame, Á.L.; Iborra, S.; Jorge, I.; et al. Priming of dendritic cells by DNA-containing extracellular vesicles from activated T cells through antigen-driven contacts. *Nat. Commun.* **2018**, *9*, 2658. [CrossRef]
64. Wang, W.; Kong, P.; Ma, G.; Li, L.; Zhu, J.; Xia, T.; Xie, H.; Zhou, W.; Wang, S. Characterization of the release and biological significance of cell-free DNA from breast cancer cell lines. *Oncotarget* **2017**, *8*, 43180–43191. [CrossRef]
65. Serpas, L.; Chan, R.W.Y.; Jiang, P.; Ni, M.; Sun, K.; Rashidfarrokhi, A.; Soni, C.; Sisirak, V.; Lee, W.-S.; Cheng, S.H.; et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 641–649. [CrossRef] [PubMed]
66. Heitzer, E.; Auinger, L.; Speicher, M.R. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends Mol. Med.* **2020**, *26*, 519–528. [CrossRef]
67. Rostami, A.; Lambie, M.; Yu, C.W.; Stambolic, V.; Waldron, J.N.; Bratman, S.V. Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics. *Cell Rep.* **2020**, *31*. [CrossRef] [PubMed]
68. Franco, R.S. Measurement of Red Cell Lifespan and Aging. *Transfus. Med. Hemotherapy* **2012**, *39*, 302–307. [CrossRef]
69. Ji, P.; Murata-Hori, M.; Lodish, H.F. Formation of mammalian erythrocytes: Chromatin condensation and enucleation. *Trends Cell Biol.* **2011**, *21*, 409–415. [CrossRef]
70. Palis, J. Primitive and definitive erythropoiesis in mammals. *Front. Physiol.* **2014**, *5*, 3. [CrossRef]
71. Moras, M.; Lefevre, S.D.; Ostuni, M.A. From Erythroblasts to Mature Red Blood Cells: Organelle Clearance in Mammals. *Front. Physiol.* **2017**, *8*. [CrossRef]
72. Keerthivasan, G.; Wickrema, A.; Crispino, J.D. Erythroblast Enucleation. *Stem Cells Int.* **2011**, *2011*, 1–9. [CrossRef] [PubMed]
73. Chasis, J.A.; Mohandas, N. Erythroblastic islands: Niches for erythropoiesis. *Blood* **2008**, *112*, 470–478. [CrossRef]
74. Gulliver, G. Observations on the size and shapes of the red corpuscles of the blood of vertebrates with drawings of them to a uniform scale and extended and revised tables of measurements. *Proc. Zool. Soc. Lond.* **1875**, 474–495.
75. Simpson, C.F.; Kling, J.M. The mechanism of denucleation in circulating erythroblasts. *J. Cell Biol.* **1967**, *35*, 237–245. [CrossRef]
76. BESSIS, M. L'îlot érythroblastique, unité fonctionnelle de la moelle osseuse. *Rev. Hematol.* **1958**, *13*, 8–11.
77. Geiduschek, J.B.; Singer, S.J. Molecular changes in the membranes of mouse erythroid cells accompanying differentiation. *Cell* **1979**, *16*, 149–163. [CrossRef]

78. McGrath, K.E.; Kingsley, P.D.; Koniski, A.D.; Porter, R.L.; Bushnell, T.P.; Palis, J. Enucleation of primitive erythroid cells generates a transient population of "pyrenocytes" in the mammalian fetus. *Blood* **2008**, *111*, 2409–2417. [CrossRef] [PubMed]
79. Yoshida, H.; Kawane, K.; Koike, M.; Mori, Y.; Uchiyama, Y.; Nagata, S. Phosphatidylserine-dependent engulfment by macrophages of nuclei from erythroid precursor cells. *Nature* **2005**, *437*, 754–758. [CrossRef] [PubMed]
80. Kawane, K. Requirement of DNase II for Definitive Erythropoiesis in the Mouse Fetal Liver. *Science* **2001**, *292*, 1546–1549. [CrossRef] [PubMed]
81. Dunaeva, M.; Buddingh', B.C.; Toes, R.E.M.; Luime, J.J.; Lubberts, E.; Pruijn, G.J.M. Decreased serum cell-free DNA levels in rheumatoid arthritis. *Autoimmun. Highlights* **2015**, *6*, 23–30. [CrossRef] [PubMed]
82. Helmig, S.; Frühbeis, C.; Krämer-Albers, E.-M.; Simon, P.; Tug, S. Release of bulk cell free DNA during physical exercise occurs independent of extracellular vesicles. *Eur. J. Appl. Physiol.* **2015**, *115*, 2271–2280. [CrossRef] [PubMed]
83. Tug, S.; Helmig, S.; Deichmann, E.R.; Schmeier-Jürchott, A.; Wagner, E.; Zimmermann, T.; Radsak, M.; Giacca, M.; Simon, P. Exercise-induced increases in cell free DNA in human plasma originate predominantly from cells of the haematopoietic lineage. *Exerc. Immunol. Rev.* **2015**, *21*, 164–173.
84. Lui, Y.Y.N.; Chik, K.-W.; Chiu, R.W.K.; Ho, C.-Y.; Lam, C.W.K.; Lo, Y.M.D. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin. Chem.* **2002**, *48*, 421–427. [CrossRef]
85. Wong, F.C.K.; Sun, K.; Jiang, P.; Cheng, Y.K.Y.; Chan, K.C.A.; Leung, T.Y.; Chiu, R.W.K.; Lo, Y.M.D. Cell-free DNA in maternal plasma and serum: A comparison of quantity, quality and tissue origin using genomic and epigenomic approaches. *Clin. Biochem.* **2016**, *49*, 1379–1386. [CrossRef]
86. Moss, J.; Magenheim, J.; Neiman, D.; Zemmour, H.; Loyfer, N.; Korach, A.; Samet, Y.; Maoz, M.; Druid, H.; Arner, P.; et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **2018**, *9*, 5068. [CrossRef]
87. Kustanovich, A.; Schwartz, R.; Peretz, T.; Grinshpun, A. Life and death of circulating cell-free DNA. *Cancer Biol. Ther.* **2019**, *20*, 1057–1067. [CrossRef]
88. Lam, W.K.J.; Gai, W.; Sun, K.; Wong, R.S.M.; Chan, R.W.Y.; Jiang, P.; Chan, N.P.H.; Hui, W.W.I.; Chan, A.W.H.; Szeto, C.-C.; et al. DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia. *Clin. Chem.* **2017**, *63*, 1614–1623. [CrossRef]
89. Brinkmann, V.; Reichard, U.; Goosmann, C.; Fauler, B.; Uhlemann, Y.; Weiss, D.S.; Weinrauch, Y.; Zychlinsky, A. Neutrophil Extracellular Traps Kill Bacteria. *Science* **2004**, *303*, 1532–1535. [CrossRef]
90. Yipp, B.G.; Kubes, P. NETosis: How vital is it? *Blood* **2013**, *122*, 2784–2794. [CrossRef]
91. Fuchs, T.A.; Abed, U.; Goosmann, C.; Hurwitz, R.; Schulze, I.; Wahn, V.; Weinrauch, Y.; Brinkmann, V.; Zychlinsky, A. Novel cell death program leads to neutrophil extracellular traps. *J. Cell Biol.* **2007**, *176*, 231–241. [CrossRef]
92. Estúa-Acosta, G.A.; Zamora-Ortiz, R.; Buentello-Volante, B.; García-Mejía, M.; Garfias, Y. Neutrophil Extracellular Traps: Current Perspectives in the Eye. *Cells* **2019**, *8*, 979. [CrossRef]
93. Branzk, N.; Papayannopoulos, V. Molecular mechanisms regulating NETosis in infection and disease. *Semin. Immunopathol.* **2013**, *35*, 513–530. [CrossRef]
94. Pilsczek, F.H.; Salina, D.; Poon, K.K.H.; Fahey, C.; Yipp, B.G.; Sibley, C.D.; Robbins, S.M.; Green, F.H.Y.; Surette, M.G.; Sugai, M.; et al. A Novel Mechanism of Rapid Nuclear Neutrophil Extracellular Trap Formation in Response to Staphylococcus aureus. *J. Immunol.* **2010**, *185*, 7413–7425. [CrossRef]
95. Hamam, H.J.; Palaniyar, N. Post-Translational Modifications in NETosis and NETs-Mediated Diseases. *Biomolecules* **2019**, *9*, 369. [CrossRef]
96. Kenny, E.F.; Herzig, A.; Krüger, R.; Muth, A.; Mondal, S.; Thompson, P.R.; Brinkmann, V.; von Bernuth, H.; Zychlinsky, A. Diverse stimuli engage different neutrophil extracellular trap pathways. *Elife* **2017**, *6*, 1–21. [CrossRef]
97. Amulic, B.; Knackstedt, S.L.; Abu Abed, U.; Deigendesch, N.; Harbort, C.J.; Caffrey, B.E.; Brinkmann, V.; Heppner, F.L.; Hinds, P.W.; Zychlinsky, A. Cell-Cycle Proteins Control Production of Neutrophil Extracellular Traps. *Dev. Cell* **2017**, *43*, 449–462.e5. [CrossRef]
98. Farrera, C.; Fadeel, B. Macrophage Clearance of Neutrophil Extracellular Traps Is a Silent Process. *J. Immunol.* **2013**, *191*, 2647–2656. [CrossRef]

99. Pisetsky, D.S. The origin and properties of extracellular DNA: From PAMP to DAMP. *Clin. Immunol.* **2012**, *144*, 32–40. [CrossRef]
100. Yangsheng, Y.; Kaihong, S. Neutrophil Extracellular Traps and Systemic Lupus Erythematosus. *J. Clin. Cell. Immunol.* **2013**, *4*. [CrossRef]
101. Leffler, J.; Martin, M.; Gullstrand, B.; Tydén, H.; Lood, C.; Truedsson, L.; Bengtsson, A.A.; Blom, A.M. Neutrophil Extracellular Traps That Are Not Degraded in Systemic Lupus Erythematosus Activate Complement Exacerbating the Disease. *J. Immunol.* **2012**, *188*, 3522–3531. [CrossRef]
102. Leffler, J.; Stojanovich, L.; Shoenfeld, Y.; Bogdanovic, G.; Hesselstrand, R.; Blom, A.M. Degradation of neutrophil extracellular traps is decreased in patients with antiphospholipid syndrome. *Clin. Exp. Rheumatol.* **2014**, *32*, 66–70.
103. Hakkim, A.; Fürnrohr, B.G.; Amann, K.; Laube, B.; Abed, U.A.; Brinkmann, V.; Herrmann, M.; Voll, R.E.; Zychlinsky, A. Impairment of neutrophil extracellular trap degradation is associated with lupus nephritis. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 9813–9818. [CrossRef] [PubMed]
104. Doke, M.; Fukamachi, H.; Morisaki, H.; Arimoto, T.; Kataoka, H.; Kuwata, H. Nucleases from Prevotella intermedia can degrade neutrophil extracellular traps. *Mol. Oral Microbiol.* **2017**, *32*, 288–300. [CrossRef]
105. Berends, E.T.M.; Horswill, A.R.; Haste, N.M.; Monestier, M.; Nizet, V.; Von Köckritz-Blickwede, M. Nuclease expression by Staphylococcus aureus facilitates escape from neutrophil extracellular traps. *J. Innate Immun.* **2010**, *2*, 576–586. [CrossRef] [PubMed]
106. de Buhr, N.; Neumann, A.; Jerjomiceva, N.; von Köckritz-Blickwede, M.; Baums, C.G. Streptococcus suis DNase SsnA contributes to degradation of neutrophil extracellular traps (NETs) and evasion of NET-mediated antimicrobial activity. *Microbiology* **2014**, *160*, 385–395. [CrossRef] [PubMed]
107. Storisteanu, D.M.L.; Pocock, J.M.; Cowburn, A.S.; Juss, J.K.; Nadesalingam, A.; Nizet, V.; Chilvers, E.R. Evasion of neutrophil extracellular traps by respiratory pathogens. *Am. J. Respir. Cell Mol. Biol.* **2017**, *56*, 423–431. [CrossRef] [PubMed]
108. Neumann, A.; Völlger, L.; Berends, E.T.M.; Molhoek, E.M.; Stapels, D.A.C.; Midon, M.; Friães, A.; Pingoud, A.; Rooijakkers, S.H.M.; Gallo, R.L.; et al. Novel role of the antimicrobial peptide LL-37 in the protection of neutrophil extracellular Traps against degradation by bacterial nucleases. *J. Innate Immun.* **2014**, *6*, 860–868. [CrossRef] [PubMed]
109. Albrengues, J.; Shields, M.A.; Ng, D.; Park, C.G.; Ambrico, A.; Poindexter, M.E.; Upadhyay, P.; Uyeminami, D.L.; Pommier, A.; Küttner, V.; et al. Neutrophil extracellular traps produced during inflammation awaken dormant cancer cells in mice. *Science* **2018**, *361*. [CrossRef]
110. Anderton, H.; Wicks, I.P.; Silke, J. Cell death in chronic inflammation: Breaking the cycle to treat rheumatic disease. *Nat. Rev. Rheumatol.* **2020**. [CrossRef]
111. Katkar, G.D.; Sundaram, M.S.; NaveenKumar, S.K.; Swethakumar, B.; Sharma, R.D.; Paul, M.; Vishalakshi, G.J.; Devaraja, S.; Girish, K.S.; Kemparaju, K. NETosis and lack of DNase activity are key factors in Echis carinatus venom-induced tissue destruction. *Nat. Commun.* **2016**, *7*. [CrossRef] [PubMed]
112. Zuo, Y.; Yalavarthi, S.; Shi, H.; Gockman, K.; Zuo, M.; Madison, J.A.; Blair, C.N.; Weber, A.; Barnes, B.J.; Egeblad, M.; et al. Neutrophil extracellular traps in COVID-19. *JCI Insight* **2020**, *9*, 1494. [CrossRef]
113. Gahan, P.B.; Stroun, M. The virtosome-a novel cytosolic informative entity and intercellular messenger. *Cell Biochem. Funct.* **2010**, *28*, 529–538. [CrossRef] [PubMed]
114. Rogers, J.C.; Boldt, D.; Kornfeld, S.; Skinner, S.A.; Valeri, C.R. Excretion of Deoxyribonucleic Acid by Lymphocytes Stimulated with Phytohemagglutinin or Antigen. *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 1685–1689. [CrossRef] [PubMed]
115. Anker, P.; Stroun, M.; Maurice, P.A. Spontaneous release of DNA by human blood lymphocytes as shown in an in vitro system. *Cancer Res.* **1975**, *35*, 2375–2382. [PubMed]
116. Anker, P.; Stroun, M.; Maurice, P.A. Spontaneous extracellular synthesis of DNA released by human blood lymphocytes. *Cancer Res.* **1976**, *36*, 2832–2839.
117. Stroun, M.; Anker, P. Nucleic acids spontaneously released by living frog auricles. *Biochem. J.* **1972**, *128*, 100–101. [CrossRef]
118. Adams, D.H.; Gahan, P.B. The DNA extruded by rat spleen cells in culture. *Int. J. Biochem.* **1983**, *15*, 547–552. [CrossRef]
119. Challen, C.; Adams, D.H. Further studies on the size and composition of the chick embryo fibroblast cytosolic DNA complex. *Int. J. Biochem.* **1986**, *18*, 423–429. [CrossRef]

120. Cataldi, S.; Viola-Magni, M. Components of the cytosolic and released virtosomes from stimulated and non-stimulated human lymphocytes. *Biochem. Biophys. Rep.* **2016**, *6*, 236–241. [CrossRef]
121. Adams, D.H.; Diaz, N.; Gahan, P.B. In vitro stimulation by tumour cell media of [3H]-thymidine incorporation by mouse spleen lymphocytes. *Cell Biochem. Funct.* **1997**, *15*, 119–126. [CrossRef]
122. García-Olmo, D.C.; Domínguez, C.; García-Arranz, M.; Anker, P.; Stroun, M.; García-Verdugo, J.M.; García-Olmo, D. Cell-free nucleic acids circulating in the plasma of colorectal cancer patients induce the oncogenic transformation of susceptible cultured cells. *Cancer Res.* **2010**, *70*, 560–567. [CrossRef]
123. Garcia-Arranz, M.; Garcia-Olmo, D.; Vega-Clemente, L.; Stroun, M.; Gahan, P.B. Non-dividing Cell Virtosomes Affect In Vitro and In Vivo Tumour Cell Replication. In *Circulating Nucleic Acids in Serum and Plasma—CNAPS IX. Advances in Experimental Medicine and Biology*; Gahan, P., Fleischhacker, M., Schmidt, B., Eds.; Springer: Cham, Switzerland, 2016; pp. 43–45.
124. Pelc, S.R. Turnover of DNA and function. *Nature* **1968**, *219*, 162–163. [CrossRef]
125. Stroun, M.; Charles, P.; Anker, P.; Pelc, S.R. Metabolic DNA in heart and skeletal muscle and in the intestine of mice. *Nature* **1967**, *216*, 716–717. [CrossRef] [PubMed]
126. Bronkhorst, A.J.; Wentzel, J.F.; Aucamp, J.; van Dyk, E.; du Plessis, L.; Pretorius, P.J. Characterization of the cell-free DNA released by cultured cancer cells. *Biochim. Biophys. Acta Mol. Cell Res.* **2016**, *1863*, 157–165. [CrossRef]
127. Aucamp, J.; Bronkhorst, A.J.; Peters, D.L.; Van Dyk, H.C.; Van der Westhuizen, F.H.; Pretorius, P.J. Kinetic analysis, size profiling, and bioenergetic association of DNA released by selected cell lines in vitro. *Cell. Mol. Life Sci.* **2017**, *74*, 2689–2707. [CrossRef] [PubMed]
128. Panagopoulou, M.; Karaglani, M.; Balgkouranidou, I.; Pantazi, C.; Kolios, G.; Kakolyris, S.; Chatzaki, E. Circulating cell-free DNA release in vitro: Kinetics, size profiling, and cancer-related gene methylation. *J. Cell. Physiol.* **2019**, *234*, 14079–14089. [CrossRef] [PubMed]
129. Thakur, B.K.; Zhang, H.; Becker, A.; Matei, I.; Huang, Y.; Costa-Silva, B.; Zheng, Y.; Hoshino, A.; Brazier, H.; Xiang, J.; et al. Double-stranded DNA in exosomes: A novel biomarker in cancer detection. *Cell Res.* **2014**, *24*, 766–769. [CrossRef] [PubMed]
130. Ronquist, G. Prostasomes are mediators of intercellular communication: From basic research to clinical implications. *J. Intern. Med.* **2012**, *271*, 400–413. [CrossRef]
131. Alcaide, M.; Cheung, M.; Hillman, J.; Rassekh, S.R.; Deyell, R.J.; Batist, G.; Karsan, A.; Wyatt, A.W.; Johnson, N.; Scott, D.W.; et al. Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR. *Sci. Rep.* **2020**, *10*. [CrossRef]
132. Heider, K.; Wan, J.C.M.; Hall, J.; Belic, J.; Boyle, S.; Hudecova, I.; Gale, D.; Cooper, W.N.; Corrie, P.G.; Brenton, J.D.; et al. Detection of ctDNA from Dried Blood Spots after DNA Size Selection. *Clin. Chem.* **2020**, *66*, 697–705. [CrossRef]
133. Garcia-Murillas, I.; Turner, N.C. Assessing HER2 amplification in plasma cfDNA. In *Methods in Molecular Biology*; Humana Press: New York, NY, USA, 2018; Volume 1768, pp. 161–172. [CrossRef]
134. Bonucci, E. Fine structure of early cartilage calcification. *J. Ultrastruct. Res.* **1967**, *20*, 33–50. [CrossRef]
135. Wolf, P. The Nature and Significance of Platelet Products in Human Plasma. *Br. J. Haematol.* **1967**, *13*, 269–288. [CrossRef] [PubMed]
136. Hessvik, N.P.; Llorente, A. Current knowledge on exosome biogenesis and release. *Cell. Mol. Life Sci.* **2018**, *75*, 193–208. [CrossRef] [PubMed]
137. Pegtel, D.M.; Gould, S.J. Exosomes. *Annu. Rev. Biochem.* **2019**, *88*, 487–514. [CrossRef]
138. Kalluri, R.; LeBleu, V.S. The biology, function, and biomedical applications of exosomes. *Science* **2020**, *367*. [CrossRef]
139. Kim, J.H.; Lee, J.; Park, J.; Gho, Y.S. Gram-negative and Gram-positive bacterial extracellular vesicles. *Semin. Cell Dev. Biol.* **2015**, *40*, 97–104. [CrossRef]
140. Margolis, L.; Sadovsky, Y. The biology of extracellular vesicles: The known unknowns. *PLoS Biol.* **2019**, *17*, e3000363. [CrossRef]
141. Raposo, G.; Stoorvogel, W. Extracellular vesicles: Exosomes, microvesicles, and friends. *J. Cell Biol.* **2013**, *200*, 373–383. [CrossRef]
142. Akers, J.C.; Gonda, D.; Kim, R.; Carter, B.S.; Chen, C.C. Biogenesis of extracellular vesicles (EV): Exosomes, microvesicles, retrovirus-like vesicles, and apoptotic bodies. *J. Neurooncol.* **2013**, *113*, 1–11. [CrossRef]

143. Doyle, L.M.; Wang, M.Z. Overview of Extracellular Vesicles, Their Origin, Composition, Purpose, and Methods for Exosome Isolation and Analysis. *Cells* **2019**, *8*, 727. [CrossRef] [PubMed]
144. Harding, C.; Heuser, J.; Stahl, P. Receptor-mediated endocytosis of transferrin and recycling of the transferrin receptor in rat reticulocytes. *J. Cell Biol.* **1983**, *97*, 329–339. [CrossRef]
145. Olsson, I.; Ronquist, G. Nucleic Acid Association to Human Prostasomes. *Arch. Androl.* **1990**, *24*, 1–10. [CrossRef]
146. Ronquist, K.G.; Ronquist, G.; Carlsson, L.; Larsson, A. Human prostasomes contain chromosomal DNA. *Prostate* **2009**, *69*, 737–743. [CrossRef] [PubMed]
147. Guescini, M.; Genedani, S.; Stocchi, V.; Agnati, L.F. Astrocytes and Glioblastoma cells release exosomes carrying mtDNA. *J. Neural Transm.* **2010**, *117*, 1–4. [CrossRef] [PubMed]
148. Dorward, D.W.; Garon, C.F.; Judd, R.C. Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. *J. Bacteriol.* **1989**, *171*, 2499–2505. [CrossRef]
149. Garon, C.F.; Dorward, D.W.; Corwin, M.D. Structural features of Borrelia burgdorferi–the Lyme disease spirochete: Silver staining for nucleic acids. *Scanning Microsc. Suppl.* **1989**, *3*, 109–115.
150. Soler, N.; Marguet, E.; Verbavatz, J.-M.; Forterre, P. Virus-like vesicles and extracellular DNA produced by hyperthermophilic archaea of the order Thermococcales. *Res. Microbiol.* **2008**, *159*, 390–399. [CrossRef]
151. Kalluri, R.; Lebleu, V.S. Discovery of double-stranded genomic DNA in circulating exosomes. *Cold Spring Harb. Symp. Quant. Biol.* **2016**, *81*, 275–280. [CrossRef]
152. Kahlert, C.; Melo, S.A.; Protopopov, A.; Tang, J.; Seth, S.; Koch, M.; Zhang, J.; Weitz, J.; Chin, L.; Futreal, A.; et al. Identification of Double-stranded Genomic DNA Spanning All Chromosomes with Mutated KRAS and p53 DNA in the Serum Exosomes of Patients with Pancreatic Cancer. *J. Biol. Chem.* **2014**, *289*, 3869–3875. [CrossRef]
153. Sansone, P.; Savini, C.; Kurelac, I.; Chang, Q.; Amato, L.B.; Strillacci, A.; Stepanova, A.; Iommarini, L.; Mastroleo, C.; Daly, L.; et al. Packaging and transfer of mitochondrial DNA via exosomes regulate escape from dormancy in hormonal therapy-resistant breast cancer. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E9066–E9075. [CrossRef]
154. Balaj, L.; Lessard, R.; Dai, L.; Cho, Y.-J.; Pomeroy, S.L.; Breakefield, X.O.; Skog, J. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nat. Commun.* **2011**, *2*, 180. [CrossRef]
155. Waldenström, A.; Gennebäck, N.; Hellman, U.; Ronquist, G. Cardiomyocyte Microvesicles Contain DNA/RNA and Convey Biological Messages to Target Cells. *PLoS ONE* **2012**, *7*, e34653. [CrossRef] [PubMed]
156. Yang, S.; Che, S.P.Y.; Kurywchak, P.; Tavormina, J.L.; Gansmo, L.B.; Correa de Sampaio, P.; Tachezy, M.; Bockhorn, M.; Gebauer, F.; Haltom, A.R.; et al. Detection of mutant KRAS and TP53 DNA in circulating exosomes from healthy individuals and patients with pancreatic cancer. *Cancer Biol. Ther.* **2017**, *18*, 158–165. [CrossRef] [PubMed]
157. Cai, J.; Han, Y.; Ren, H.; Chen, C.; He, D.; Zhou, L.; Eisner, G.M.; Asico, L.D.; Jose, P.A.; Zeng, C. Extracellular vesicle-mediated transfer of donor genomic DNA to recipient cells is a novel mechanism for genetic influence between cells. *J. Mol. Cell Biol.* **2013**, *5*, 227–238. [CrossRef] [PubMed]
158. Cai, J.; Wu, G.; Jose, P.A.; Zeng, C. Functional transferred DNA within extracellular vesicles. *Exp. Cell Res.* **2016**, *349*, 179–183. [CrossRef]
159. Sharma, A.; Johnson, A. Exosome DNA: Critical regulator of tumor immunity and a diagnostic biomarker. *J. Cell. Physiol.* **2020**, *235*, 1921–1932. [CrossRef]
160. Bitto, N.J.; Chapman, R.; Pidot, S.; Costin, A.; Lo, C.; Choi, J.; D'Cruze, T.; Reynolds, E.C.; Dashper, S.G.; Turnbull, L.; et al. Bacterial membrane vesicles transport their DNA cargo into host cells. *Sci. Rep.* **2017**, *7*, 7072. [CrossRef]
161. Fischer, S.; Cornils, K.; Speiseder, T.; Badbaran, A.; Reimer, R.; Indenbirken, D.; Grundhoff, A.; Brunswig-Spickenheier, B.; Alawi, M.; Lange, C. Indication of Horizontal DNA Gene Transfer by Extracellular Vesicles. *PLoS ONE* **2016**, *11*, e0163665. [CrossRef]
162. Khier, S.; Lohan, L. Kinetics of circulating cell-free DNA for biomedical applications: Critical appraisal of the literature. *Future Sci. OA* **2018**, *4*, FSO295. [CrossRef]
163. Lázaro-Ibáñez, E.; Lässer, C.; Shelke, G.V.; Crescitelli, R.; Jang, S.C.; Cvjetkovic, A.; García-Rodríguez, A.; Lötvall, J. DNA analysis of low- and high-density fractions defines heterogeneous subpopulations of small extracellular vesicles based on their DNA cargo and topology. *J. Extracell. Vesicles* **2019**, *8*, 1656993. [CrossRef] [PubMed]

164. Takahashi, A.; Okada, R.; Nagao, K.; Kawamata, Y.; Hanyu, A.; Yoshimoto, S.; Takasugi, M.; Watanabe, S.; Kanemaki, M.T.; Obuse, C.; et al. Exosomes maintain cellular homeostasis by excreting harmful DNA from cells. *Nat. Commun.* **2017**, *8*. [CrossRef] [PubMed]
165. Baixauli, F.; López-Otín, C.; Mittelbrunn, M. Exosomes and autophagy: Coordinated mechanisms for the maintenance of cellular fitness. *Front. Immunol.* **2014**, *5*. [CrossRef] [PubMed]
166. Desdín-Micó, G.; Mittelbrunn, M. Role of exosomes in the protection of cellular homeostasis. *Cell Adhes. Migr.* **2017**, *11*, 127–134. [CrossRef]
167. Brahmer, A.; Neuberger, E.; Esch-Heisser, L.; Haller, N.; Jorgensen, M.M.; Baek, R.; Möbius, W.; Simon, P.; Krämer-Albers, E.M. Platelets, endothelial cells and leukocytes contribute to the exercise-triggered release of extracellular vesicles into the circulation. *J. Extracell. Vesicles* **2019**, *8*. [CrossRef]
168. Bronkhorst, A.J.; Wentzel, J.F.; Ungerer, V.; Peters, D.L.; Aucamp, J.; de Villiers, E.P.; Holdenrieder, S.; Pretorius, P.J. Sequence analysis of cell-free DNA derived from cultured human bone osteosarcoma (143B) cells. *Tumor Biol.* **2018**, *40*. [CrossRef]
169. Grabuschnig, S.; Soh, J.; Heidinger, P.; Bachler, T.; Hirschböck, E.; Rosales Rodriguez, I.; Schwendenwein, D.; Sensen, C.W. Circulating cell-free DNA is predominantly composed of retrotransposable elements and non-telomeric satellite DNA. *J. Biotechnol.* **2020**, *313*, 48–56. [CrossRef]
170. Jefford, C.E.; Irminger-Finger, I. Mechanisms of chromosome instability in cancers. *Crit. Rev. Oncol. Hematol.* **2006**, *59*, 1–14. [CrossRef]
171. Westhorpe, F.G.; Straight, A.F. Functions of the centromere and kinetochore in chromosome segregation. *Curr. Opin. Cell Biol.* **2013**, *25*, 334–340. [CrossRef]
172. Cimini, D.; Degrassi, F. Aneuploidy: A matter of bad connections. *Trends Cell Biol.* **2005**, *15*, 442–451. [CrossRef]
173. Cimini, D.; Moree, B.; Canman, J.C.; Salmon, E.D. Merotelic kinetochore orientation occurs frequently during early mitosis in mammalian tissue cells and error correction is achieved by two different mechanisms. *J. Cell Sci.* **2003**, *116*, 4213–4225. [CrossRef] [PubMed]
174. Lampson, M.A.; Renduchitala, K.; Khodjakov, A.; Kapoor, T.M. Correcting improper chromosome-spindle attachments during cell division. *Nat. Cell Biol.* **2004**, *6*, 232–237. [CrossRef] [PubMed]
175. Cimini, D.; Howell, B.; Maddox, P.; Khodjakov, A.; Degrassi, F.; Salmon, E.D. Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J. Cell Biol.* **2001**, *153*, 517–527. [CrossRef] [PubMed]
176. Sears, D.A.; Udden, M.M. Howell-Jolly bodies: A brief historical review. *Am. J. Med. Sci.* **2012**, *343*, 407–409. [CrossRef] [PubMed]
177. Holland, N.; Bolognesi, C.; Kirsch-Volders, M.; Bonassi, S.; Zeiger, E.; Knasmueller, S.; Fenech, M. The micronucleus assay in human buccal cells as a tool for biomonitoring DNA damage: The HUMN project perspective on current status and knowledge gaps. *Mutat. Res.* **2008**, *659*, 93–108. [CrossRef]
178. Ford, J.H.; Schultz, C.J.; Correll, A.T. Chromosome elimination in micronuclei: A common cause of hypoploidy. *Am. J. Hum. Genet.* **1988**, *43*, 733–740.
179. Lindholm, C.; Norppa, H.; Hayashi, M.; Sorsa, M. Induction of micronuclei and anaphase aberrations by cytochalasin B in human lymphocyte cultures. *Mutat. Res.* **1991**, *260*, 369–375. [CrossRef]
180. Ford, J.H.; Correll, A.T. Chromosome errors at mitotic anaphase. *Genome* **1992**, *35*, 702–705. [CrossRef]
181. Catalán, J.; Falck, G.C.; Norppa, H. The X chromosome frequently lags behind in female lymphocyte anaphase. *Am. J. Hum. Genet.* **2000**, *66*, 687–691. [CrossRef]
182. Norppa, H.; Falck, G.C.-M. What do human micronuclei contain? *Mutagenesis* **2003**, *18*, 221–233. [CrossRef]
183. Bailey, S.M.; Murnane, J.P. Telomeres, chromosome instability and cancer. *Nucleic Acids Res.* **2006**, *34*, 2408–2417. [CrossRef] [PubMed]
184. Gisselsson, D.; Björk, J.; Höglund, M.; Mertens, F.; Dal Cin, P.; Akerman, M.; Mandahl, N. Abnormal nuclear shape in solid tumors reflects mitotic instability. *Am. J. Pathol.* **2001**, *158*, 199–206. [CrossRef]
185. Gisselsson, D.; Jonson, T.; Petersén, A.; Strömbeck, B.; Dal Cin, P.; Höglund, M.; Mitelman, F.; Mertens, F.; Mandahl, N. Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 12683–12688. [CrossRef]

186. Meeker, A.K.; Hicks, J.L.; Iacobuzio-Donahue, C.A.; Montgomery, E.A.; Westra, W.H.; Chan, T.Y.; Ronnett, B.M.; De Marzo, A.M. Telomere length abnormalities occur early in the initiation of epithelial carcinogenesis. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2004**, *10*, 3317–3326. [CrossRef] [PubMed]
187. Bryan, T.M.; Englezou, A.; Dunham, M.A.; Reddel, R.R. Telomere length dynamics in telomerase-positive immortal human cell populations. *Exp. Cell Res.* **1998**, *239*, 370–378. [CrossRef]
188. Maciejowski, J.; de Lange, T. Telomeres in cancer: Tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 175–186. [CrossRef]
189. Hemann, M.T.; Strong, M.A.; Hao, L.Y.; Greider, C.W. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell* **2001**, *107*, 67–77. [CrossRef]
190. Hedges, D.J.; Deininger, P.L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat. Res.* **2007**, *616*, 46–59. [CrossRef]
191. Gasior, S.L.; Wakeman, T.P.; Xu, B.; Deininger, P.L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **2006**, *357*, 1383–1393. [CrossRef]
192. Ji, W.; Hernandez, R.; Zhang, X.Y.; Qu, G.Z.; Frady, A.; Varela, M.; Ehrlich, M. DNA demethylation and pericentromeric rearrangements of chromosome 1. *Mutat. Res.* **1997**, *379*, 33–41. [CrossRef]
193. Medvedeva, N.G.; Panyutin, I.V.; Panyutin, I.G.; Neumann, R.D. Phosphorylation of histone H2AX in radiation-induced micronuclei. *Radiat. Res.* **2007**, *168*, 493–498. [CrossRef] [PubMed]
194. Stewénius, Y.; Gorunova, L.; Jonson, T.; Larsson, N.; Höglund, M.; Mandahl, N.; Mertens, F.; Mitelman, F.; Gisselsson, D. Structural and numerical chromosome changes in colon cancer develop through telomere-mediated anaphase bridges, not through mitotic multipolarity. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 5541–5546. [CrossRef] [PubMed]
195. Cornforth, M.N.; Goodwin, E.H. Transmission of radiation-induced acentric chromosomal fragments to micronuclei in normal human fibroblasts. *Radiat. Res.* **1991**, *126*, 210–217. [CrossRef] [PubMed]
196. Warburton, P.E.; Greig, G.M.; Haaf, T.; Willard, H.F. PCR amplification of chromosome-specific alpha satellite DNA: Definition of centromeric STS markers and polymorphic analysis. *Genomics* **1991**, *11*, 324–333. [CrossRef]
197. Fenech, M. Folate, DNA damage and the aging brain. *Mech. Ageing Dev.* **2010**, *131*, 236–241. [CrossRef]
198. Lindberg, H.K.; Wang, X.; Järventaus, H.; Falck, G.C.-M.; Norppa, H.; Fenech, M. Origin of nuclear buds and micronuclei in normal and folate-deprived human lymphocytes. *Mutat. Res.* **2007**, *617*, 33–45. [CrossRef]
199. Stimpson, K.M.; Matheny, J.E.; Sullivan, B.A. Dicentric chromosomes: Unique models to study centromere function and inactivation. *Chromosom. Res. Int. J. Mol. Supramol. Evol. Asp. Chromosom. Biol.* **2012**, *20*, 595–605. [CrossRef]
200. Mackinnon, R.N.; Campbell, L.J. The role of dicentric chromosome formation and secondary centromere deletion in the evolution of myeloid malignancy. *Genet. Res. Int.* **2011**, *2011*, 643628. [CrossRef]
201. McClintock, B. The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. *Proc. Natl. Acad. Sci. USA* **1942**, *28*, 458–463. [CrossRef]
202. Fenech, M.; Kirsch-Volders, M.; Natarajan, A.T.; Surralles, J.; Crott, J.W.; Parry, J.; Norppa, H.; Eastmond, D.A.; Tucker, J.D.; Thomas, P. Molecular mechanisms of micronucleus, nucleoplasmic bridge and nuclear bud formation in mammalian and human cells. *Mutagenesis* **2011**, *26*, 125–132. [CrossRef]
203. Shimizu, N.; Shimura, T.; Tanaka, T. Selective elimination of acentric double minutes from cancer cells through the extrusion of micronuclei. *Mutat. Res.* **2000**, *448*, 81–90. [CrossRef]
204. Shimizu, N.; Shingaki, K.; Kaneko-Sasaguri, Y.; Hashizume, T.; Kanda, T. When, where and how the bridge breaks: Anaphase bridge breakage plays a crucial role in gene amplification and HSR generation. *Exp. Cell Res.* **2005**, *302*, 233–243. [CrossRef] [PubMed]
205. Shimizu, N. Molecular mechanisms of the origin of micronuclei from extrachromosomal elements. *Mutagenesis* **2011**, *26*, 119–123. [CrossRef]
206. Fenech, M. Cytokinesis-block micronucleus assay evolves into a "cytome" assay of chromosomal instability, mitotic dysfunction and cell death. *Mutat. Res.* **2006**, *600*, 58–66. [CrossRef]
207. Shimizu, N.; Itoh, N.; Utiyama, H.; Wahl, G.M. Selective entrapment of extrachromosomally amplified DNA by nuclear budding and micronucleation during S phase. *J. Cell Biol.* **1998**, *140*, 1307–1320. [CrossRef]
208. Kisurina-Evgenieva, O.P.; Sutiagina, O.I.; Onishchenko, G.E. Biogenesis of Micronuclei. *Biochemistry* **2016**, *81*, 453–464. [CrossRef]

209. Utani, K.; Kohno, Y.; Okamoto, A.; Shimizu, N. Emergence of micronuclei and their effects on the fate of cells under replication stress. *PLoS ONE* **2010**, *5*, e10089. [CrossRef] [PubMed]
210. Utani, K.; Okamoto, A.; Shimizu, N. Generation of micronuclei during interphase by coupling between cytoplasmic membrane blebbing and nuclear budding. *PLoS ONE* **2011**, *6*, e27233. [CrossRef] [PubMed]
211. Sin, S.T.K.; Jiang, P.; Deng, J.; Ji, L.; Cheng, S.H.; Dutta, A.; Leung, T.Y.; Chan, K.C.A.; Chiu, R.W.K.; Lo, Y.M.D. Identification and characterization of extrachromosomal circular DNA in maternal plasma. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1658–1665. [CrossRef] [PubMed]
212. Kumar, P.; Dillon, L.W.; Shibata, Y.; Jazaeri, A.A.; Jones, D.R.; Dutta, A. Normal and Cancerous Tissues Release Extrachromosomal Circular DNA (eccDNA) into the Circulation. *Mol. Cancer Res.* **2017**, *15*, 1197–1205. [CrossRef]
213. Zhu, J.; Zhang, F.; Du, M.; Zhang, P.; Fu, S.; Wang, L. Molecular characterization of cell-free eccDNAs in human plasma. *Sci. Rep.* **2017**, *7*, 10968. [CrossRef]
214. Pampalona, J.; Soler, D.; Genescà, A.; Tusell, L. Telomere dysfunction and chromosome structure modulate the contribution of individual chromosomes in abnormal nuclear morphologies. *Mutat. Res.* **2010**, *683*, 16–22. [CrossRef] [PubMed]
215. Ehrlich, M.; Jackson, K.; Weemaes, C. Immunodeficiency, centromeric region instability, facial anomalies syndrome (ICF). *Orphanet J. Rare Dis.* **2006**, *1*, 2. [CrossRef]
216. Xu, G.L.; Bestor, T.H.; Bourc'his, D.; Hsieh, C.L.; Tommerup, N.; Bugge, M.; Hulten, M.; Qu, X.; Russo, J.J.; Viegas-Péquignot, E. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **1999**, *402*, 187–191. [CrossRef]
217. Qu, G.Z.; Grundy, P.E.; Narayan, A.; Ehrlich, M. Frequent hypomethylation in Wilms tumors of pericentromeric DNA in chromosomes 1 and 16. *Cancer Genet. Cytogenet.* **1999**, *109*, 34–39. [CrossRef]
218. Qu, G.; Dubeau, L.; Narayan, A.; Yu, M.C.; Ehrlich, M. Satellite DNA hypomethylation vs. overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. *Mutat. Res.* **1999**, *423*, 91–101. [CrossRef]
219. Narayan, A.; Ji, W.; Zhang, X.Y.; Marrogi, A.; Graff, J.R.; Baylin, S.B.; Ehrlich, M. Hypomethylation of pericentromeric DNA in breast adenocarcinomas. *Int. J. Cancer* **1998**, *77*, 833–838. [CrossRef]
220. Fauth, E.; Scherthan, H.; Zankl, H. Chromosome painting reveals specific patterns of chromosome occurrence in mitomycin C- and diethylstilboestrol-induced micronuclei. *Mutagenesis* **2000**, *15*, 459–467. [CrossRef]
221. Fauth, E.; Scherthan, H.; Zankl, H. Frequencies of occurrence of all human chromosomes in micronuclei from normal and 5-azacytidine-treated lymphocytes as revealed by chromosome painting. *Mutagenesis* **1998**, *13*, 235–241. [CrossRef]
222. Fauth, E.; Zankl, H. Comparison of spontaneous and idoxuridine-induced micronuclei by chromosome painting. *Mutat. Res.* **1999**, *440*, 147–156. [CrossRef]
223. Cimini, D.; Tanzarella, C.; Degrassi, F. Effects of 5-azacytidine on the centromeric region of human fibroblasts studied by CREST staining and in situ hybridization on cytokinesis-blocked cells. *Cytogenet. Cell Genet.* **1996**, *72*, 219–224. [CrossRef] [PubMed]
224. Guttenbach, M.; Schmid, M. Exclusion of specific human chromosomes into micronuclei by 5-azacytidine treatment of lymphocyte cultures. *Exp. Cell Res.* **1994**, *211*, 127–132. [CrossRef]
225. Beck, J.; Urnovitz, H.B.; Riggert, J.; Clerici, M.; Schütz, E. Profile of the Circulating DNA in apparently healthy individuals. *Clin. Chem.* **2009**, *55*, 730–738. [CrossRef] [PubMed]
226. Mittra, I.; Kumar Khare, N.; Venkata Raghuram, G.; Chaubal, R.; Khambatti, F.; Gupta, D.; Gaikwad, A.; Prasannan, P.; Singh, A.; Iyer, A.; et al. Circulating nucleic acids damage DNA of healthy cells by integrating into their genomes. *J. Biosci.* **2018**, *40*, 91–111. [CrossRef] [PubMed]
227. Podgornaya, O.I.; Vasilyeva, I.N.; Bespalov, V.G. Heterochromatic Tandem Repeats in the Extracellular DNA. *Adv. Exp. Med. Biol.* **2016**, *924*, 85–89. [CrossRef]
228. Snyder, G.K.; Sheafor, B.A. Red Blood Cells: Centerpiece in the Evolution of the Vertebrate Circulatory System. *Am. Zool.* **1999**, *39*, 189–198. [CrossRef]
229. Föller, M.; Huber, S.M.; Lang, F. Erythrocyte programmed cell death. *IUBMB Life* **2008**, *60*, 661–668. [CrossRef]
230. Thiam, H.R.; Wong, S.L.; Wagner, D.D.; Waterman, C.M. Cellular Mechanisms of NETosis. *Annu. Rev. Cell Dev. Biol.* **2020**, *36*. [CrossRef]
231. ADAMS, D.H.; GAHAN, P.B. Stimulated and Non-stimulated Rat Spleen Cells Release Different DNA-Complexes. *Differentiation* **1982**, *22*, 47–52. [CrossRef]

232. McIntosh, A.A.G.; Adams, D.H. Further studies on the extrusion of cytosol macromolecules by cultured chick embryo fibroblast cells. *Int. J. Biochem.* **1985**, *17*, 147–153. [CrossRef]
233. Khandjian, E.W.; Turian, G. Release of RNA-DNA-protein complex during differentiation of the water mould Allomyces arbuscula. *Cell Differ.* **1976**, *5*, 171–188. [CrossRef]
234. Elzanowska, J.; Semira, C.; Costa-Silva, B. DNA in extracellular vesicles: Biological and clinical aspects. *Mol. Oncol.* **2020**. [CrossRef]
235. Ermakov, A.V.; Konkova, M.S.; Kostyuk, S.V.; Egolina, N.A.; Efremova, L.V.; Veiko, N.N. Oxidative stress as a significant factor for development of an adaptive response in irradiated and nonirradiated human lymphocytes after inducing the bystander effect by low-dose X-radiation. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **2009**, *669*, 155–161. [CrossRef] [PubMed]
236. Ermakov, A.V.; Konkova, M.S.; Kostyuk, S.V.; Smirnova, T.D.; Malinovskaya, E.M.; Efremova, L.V.; Veiko, N.N. An extracellular DNA mediated bystander effect produced from low dose irradiated endothelial cells. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **2011**, *712*, 1–10. [CrossRef] [PubMed]
237. Konkova, M.S.; Kaliyanov, A.A.; Sergeeva, V.A.; Abramova, M.S.; Kostyuk, S.V. Oxidized Cell-Free DNA Is a Factor of Stress Signaling in Radiation-Induced Bystander Effects in Different Types of Human Cells. *Int. J. Genom.* **2019**, *2019*. [CrossRef] [PubMed]
238. Ariyoshi, K.; Miura, T.; Kasai, K.; Fujishima, Y.; Nakata, A.; Yoshida, M. Radiation-Induced Bystander Effect is Mediated by Mitochondrial DNA in Exosome-Like Vesicles. *Sci. Rep.* **2019**, *9*. [CrossRef]
239. Brown, P. The Cobas®EGFR Mutation Test v2 assay. *Future Oncol.* **2016**, *12*, 451–452. [CrossRef]
240. André, F.; Ciruelos, E.; Rubovszky, G.; Campone, M.; Loibl, S.; Rugo, H.S.; Iwata, H.; Conte, P.; Mayer, I.A.; Kaufman, B.; et al. Alpelisib for PIK3CA-mutated, hormone receptor-positive advanced breast cancer. *N. Engl. J. Med.* **2019**, *380*, 1929–1940. [CrossRef]
241. Warren, J.D.; Xiong, W.; Bunker, A.M.; Vaughn, C.P.; Furtado, L.V.; Roberts, W.L.; Fang, J.C.; Samowitz, W.S.; Heichman, K.A. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med.* **2011**, *9*. [CrossRef]
242. Keller, L.; Belloum, Y.; Wikman, H.; Pantel, K. Clinical relevance of blood-based ctDNA analysis: Mutation detection and beyond. *Br. J. Cancer* **2020**. [CrossRef]
243. Chen, M.; Zhao, H. Next-generation sequencing in liquid biopsy: Cancer screening and early detection. *Hum. Genom.* **2019**, *13*, 34. [CrossRef] [PubMed]
244. Razavi, P.; Li, B.T.; Brown, D.N.; Jung, B.; Hubbell, E.; Shen, R.; Abida, W.; Juluru, K.; De Bruijn, I.; Hou, C.; et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* **2019**, *25*, 1928–1937. [CrossRef] [PubMed]
245. Heitzer, E.; Haque, I.S.; Roberts, C.E.S.; Speicher, M.R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* **2019**, *20*, 71–88. [CrossRef]
246. Koffler, D.; Agnello, V.; Winchester, R.; Kunkel, H.G. The occurrence of single-stranded DNA in the serum of patients with systemic lupus erythematosus and other diseases. *J. Clin. Investig.* **1973**, *52*, 198–204. [CrossRef] [PubMed]
247. Beiter, T.; Fragasso, A.; Hudemann, J.; Nieß, A.M.; Simon, P. Short-Term Treadmill Running as a Model for Studying Cell-Free DNA Kinetics In Vivo. *Clin. Chem.* **2011**, *57*, 633–636. [CrossRef]
248. Ullrich, E.; Heidinger, P.; Soh, J.; Villanova, L.; Grabuschnig, S.; Bachler, T.; Hirschböck, E.; Sánchez-Heredero, S.; Ford, B.; Sensen, M.; et al. Evaluation of host-based molecular markers for the early detection of human sepsis. *J. Biotechnol.* **2020**, *310*, 80–88. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 4

The composition of cell-free DNA in blood serum and plasma

4.1 Prologue

Around summer 2018 we still experienced several unsolved problems in our main project on diagnostic cfDNA biomarkers (see Chapter 5). Although we found that the cfDNA sequencing data was very non-uniformly distributed over the genome and this distribution was apparently altered in connection to disease conditions, the utilization of eventual signals via qPCR still proved to be challenging. While we tried to find a suitable method for qPCR data normalization, I also was working on the refinement of sequencing data analysis and the improvement of biomarker prediction. Initially, we wanted to avoid using marker sequences from repeat regions. At this point we did not consider the cfDNA from repeat regions to be informative regarding physiological host conditions. Additionally, the specificity of PCR primers for distinct repeat regions may not be ensured due to numerous highly similar regions on the genome. For this reason, I used repeat annotations from the UCSC Repeat Masker database [79] in my approach for the prediction of biomarkers in order to implement an optional masking of known repeat regions. Since I also included and displayed these repeat annotations in coverage plots, I noticed that a lot of the more strongly covered regions are directly centered over repeat elements of the Alu and long interspersed nuclear element 1 (LINE-1) family. This incited the idea that the presence of DNA fragments from these repeat elements is somehow related to the activity of retro-transposable elements (RTEs), which might even be linked to gene expression. I made the investigation of this hypothesis into my personal side project, which was based on the sequencing data from our main projects on human sepsis and Johne's Disease. This is where implementing my own analysis framework proved to be advantageous. It allowed me to quickly implement a prototype for analyzing the composition of cfDNA in regard to annotated repeat families, which showed that DNA from RTEs was indeed overrepresented. But RTEs were not the only repeat elements with an increased fraction. Following these initial results, we decided to perform an in-depth analysis of cfDNA sequencing data compositions, which is described within this chapter.

4.2 Article

The article was accepted for publication in the Elsevier Journal of Biotechnology on March 4th 2020 and published on March 9th 2020.
DOI: 10.1016/j.jbiotec.2020.03.002

4.3 Author rights and permissions

Elsevier grants authors the rights to share and use their works for personal, scholarly or institutional purposes and explicitly permits inclusion of articles in theses and dissertations under the condition that a DOI Link to the original version on Science Direct is specified.
https://www.elsevier.com/about/policies/copyright

Journal of Biotechnology 313 (2020) 48–56

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Circulating cell-free DNA is predominantly composed of retrotransposable elements and non-telomeric satellite DNA

Stefan Grabuschnig[a], Jung Soh[b], Petra Heidinger[c], Thorsten Bachler[c], Elisabeth Hirschböck[d], Ingund Rosales Rodriguez[d], Daniel Schwendenwein[c], Christoph W. Sensen[a,b,d,e,*]

[a] Graz University of Technology, Institute of Computational Biotechnology, Petersgasse 14(V), 8010, Graz, Austria
[b] CNA Diagnostics Inc., Suite 300, 4838 Richard Rd SW, Calgary, AB T3E 6L1, Canada
[c] Acib GmbH, Petersgasse 14, 8010, Graz, Austria
[d] CNA Diagnostics GmbH, Parkring 18, 8074, Grambach, Austria
[e] BioTechMed Graz, Mozartgasse 12/II, 8010, Graz, Austria

ARTICLE INFO

Keywords:
Circulating nucleic acids
Cell-free DNA
Active release
Retrotransposable elements
Satellite DNA
Exosomes

ABSTRACT

Circulating cell-free DNAs (cfDNAs) are DNA fragments which can be isolated from mammalian blood serum or plasma. In order to gain deeper insight into their origin(s), we have characterized the composition of human and cattle cfDNA via large-scale analyses of high-throughput sequencing data. We observed significant differences between the composition of cfDNA in serum/plasma and the corresponding DNA sequence composition of the human genome. Retrotransposable elements and non-telomeric satellite DNA were particularly overrepresented in the cfDNA population, while telomeric satellite DNA was underrepresented. This was consistently observed for human plasma, bovine serum and for the supernatant of human cancer cell cultures. Our results suggest that reverse transcription of retrotransposable elements and secondary-structure formation during the replication of satellite DNA are contributing to the composition of the cfDNA molecules in the mammalian blood stream. We believe that our work is an important step towards the understanding of the biogenesis of cfDNAs and thus may also facilitate the future exploitation of their diagnostic potential.

1. Introduction

Circulating nucleic acids (CNAs) in the cell-free fraction of blood (Mandel and Metais, 1948) (i.e., serum or plasma) are increasingly receiving attention as sources of biomarkers (Suraj et al., 2016; Jiang and Lo, 2016; Wan et al., 2017) for molecular diagnostics. Although the term CNA also encompasses various species of RNAs (Danielson et al., 2017; Umu et al., 2018), it usually refers to double-stranded DNA molecules (Stroun et al., 1987; Rumore and Steinman, 1990). In order to avoid ambiguities, we are using cfDNA as an acronym for circulating cell-free DNA in the following text. Approaches to gain information about host conditions from cfDNA range from length profiling (Giacona et al., 1998) over screening for disease-specific genetic alterations (Chan et al., 2013a) to the characterization of epigenetic modifications (Chan et al., 2013b). While the potential of cfDNA biomarkers for cancer diagnostics (Stroun et al., 1989; Fleischhacker and Schmidt, 2007), prenatal diagnostics (Lo et al., 2007; Breveglieri et al., 2019) and other fields has already been studied intensely, the biological mechanisms behind their occurrence and possible biological role(s) are currently not well-understood. Passive cfDNA release mechanisms, such as apoptosis and necrosis, are thought to contribute a major fraction to the total quantity of cfDNA (Stroun et al., 2001a; Holdenrieder et al., 2005), but also spontaneous release of newly synthesized DNA molecules was observed for tumour cells (Stroun et al., 2001a; Bronkhorst et al., 2016) and likewise for healthy cells (Suraj et al., 2016; Anker et al., 1975). While circulating DNA fragments have been described as being mostly shorter than 200 base pairs (bp) in length (McCoubrey-Hoyer et al., 1984), they have also been reported to form discrete bands on agarose gels, which correspond to a 150–170 bp ladder, ranging from around 150 bp to approximately 1000 bp (Rumore and Steinman, 1990; Giacona et al., 1998). This ladder is characteristic for oligonucleosomal DNA resulting from endonucleolytic cleavage during apoptosis (Holdenrieder et al., 2001; Majno and Joris, 1995; Arends et al., 1990). It is therefore believed that fragments originating from apoptotic cells are the result of nucleosome occupation (Snyder et al., 2016), where DNA wound around nucleosomes is thought to be protected from digestion due to being bound to histones. The active release of cfDNA from cells of various tissues was observed in the form of exosomes

[*] Corresponding author at: Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14/V, 8010, Graz, Austria.
E-mail address: csensen@tugraz.at (C.W. Sensen).

https://doi.org/10.1016/j.jbiotec.2020.03.002
Received 20 December 2019; Received in revised form 26 February 2020; Accepted 4 March 2020
Available online 09 March 2020

(Kahlert et al., 2014; Kalluri and LeBleu, 2016; Thakur et al., 2014), which are excreted via plasma membrane fusion of multivesicular bodies (Mayers and Audhya, 2012; Lee et al., 2012), or microvesicles (Balaj et al., 2011), which emerge directly from budding of the plasma membrane (Lee et al., 2012; Cocucci et al., 2009), as well as in the form of glycolipoprotein complexes (Gahan and Stroun, 2010). A multitude of mechanisms seems to contribute to the cfDNA concentration in human, or more general, mammalian blood plasma or serum (Thierry et al., 2016; Aucamp et al., 2018). Understanding the origin of these cfDNA molecules may be crucial for diagnostical approaches to assess physiological conditions via the cfDNA population.

In a previous study of the value of cfDNA biomarkers for the diagnosis of human sepsis (Ullrich et al., 2020) we have discovered that cfDNA molecules were frequently related to repetitive elements in the human genome, with long interspersed nuclear element 1 (LINE-1, or in short L1) and Alu repeats being particularly prominent. Takahashi et al. 2017 (Takahashi et al., 2017) showed that the excretion of cytosolic DNA fragments via exosomes is an important physiological process for cellular homeostasis. The inhibition of exosome secretion leads to the accumulation of cytoplasmic DNA, which ultimately results in senescence or apoptosis.

It was our hypothesis that the activity of retrotransposable elements (RTEs) may lead to the generation of cytosolic DNA fragments, which are subsequently excreted into the extracellular space. We have analysed high-throughput DNA sequencing data of blood plasma samples from sepsis patients and a control group of healthy individuals and also blood serum samples from healthy cattle, characterizing their DNA composition, especially with regard to repeat families. Additionally, the cell-free supernatant of cancer cell cultures was analysed to investigate the potential differences in the composition of actively released cfDNA. Since these cell lines are immortal in vitro, under cell culture conditions the supernatant was expected to contain a significantly lower fraction of cell-death related DNA in comparison to actively secreted DNA.

2. Materials and methods

2.1. Sample sets

Prof. Dr. Robert Krause´s team at the Medical University of Graz, Austria collected 219 human plasma samples. The team of Prof. Dr. Lorenz Khol (University of Veterinary Medicine, Vienna, Austria) collected 123 bovine serum samples. Dr. Amin El-Heliebi (Medical University of Graz, Austria) provided six samples of cancer cell culture supernatant from placenta choriocarcinoma (BeWo (Pattillo and Gey, 1968)) and prostate cancer (VCaP (Korenchuk et al., 2001)), which are immortal in vitro (Korenchuk et al., 2001; Millan et al., 2014).

The set of plasma samples from healthy human individuals consisted of 50 samples from 50 individuals (27 females, 23 males). The age of the individuals was 43 years on average and ranged from 19 to 86 years (information about age was missing for two individuals). The set of plasma samples from sepsis patients consisted of 169 samples from 76 individuals (32 females, 37 males, seven unknown). The age of the sepsis patients was 77 years on average and ranged from 18 to 96 years (Information about age was missing for eight patients). The set of 123 bovine serum samples was collected from 87 female individuals, where information about age was not provided. An overview of all samples is shown in Table 1.

2.2. Nucleic acid extraction and amplification of cfDNA

For nucleic acid extraction, the High Pure Viral Nucleic Acid Kit (Roche Applied Science) was used according to the manufacturer's protocol, preserving the full spectrum of circulating DNAs regardless of their containment in exosomes, (micro-) vesicles, apoptotic bodies or being free-floating in the blood. Since serum or plasma preparation and freezing did not usually occur immediately after the blood draw, the concentration of more unstable species of cfDNA may have been reduced. The GenomePlex® Single Cell Whole Genome Amplification Kit (Sigma-Aldrich) was used for the amplification of total nucleic acids according to the manufacturer's protocol. The amplified DNA was purified with the GenElute™ PCR Clean-Up Kit (Sigma-Aldrich) and sent for sequencing to SEQ-IT GmbH (Kaiserslautern, Germany).

Table 1
Overview of human blood plasma samples, bovine serum samples and cultured cell lines used in this study.

Set	Status		Individuals	Samples	
Bovine *Bos taurus domesticus*	healthy pure female		87	123	
Human *Homo sapiens sapiens*	healthy		50	50	
	sepsis	*Escherichia coli*	21	169	42
		Staphylococcus aureus	19		44
		Staphylococcus epidermidis	1		5
		invasive *Candida* species	35		78
Human cancer cell cultures	placenta choriocarcinoma BeWo1425		–	2	
	placenta choriocarcinoma BeWo1426			2	
	prostate cancer VCaP			2	

2.3. DNA concentration

DNA concentration was measured before and after amplification with a Qubit 3.0 Fluorimeter (Thermo Fisher Scientific) according to the manufacturer's protocol. Measurements were performed for a subset of 10 randomly selected samples of healthy human individuals, sepsis patients, bovina and cell culture supernatants. The DNA concentration of non-amplified cfDNA was not determined, as DNA levels were below the detection limit of the device. After amplification, the DNA concentration was 24.4 ± 4.3 ng/μl (mean and standard deviation) for the subset from plasma samples from healthy human individuals, 27.2 ± 6.7 ng/μl for the subset of plasma samples from sepsis patients, 30.8 ± 9.8 ng/μl for the subset of bovine serum samples and 12.8 ± 2.8 ng/μl for cancer cell culture supernatants.

2.4. DNA sequencing

The cfDNA samples were sequenced by SEQ-IT GmbH (Kaiserslautern, Germany) on an Illumina NextSeq sequencer. Paired-end sequencing was performed, producing 150-bp long reads from each end. The sequencing depth was approximately five million reads per sample in each sequencing direction, for a total of approximately 10 million reads per sample.

2.5. Pre-processing and mapping

Remnants of Nextera Transposase adapter sequences were removed from the sequencing data using Trimmomatic (Bolger et al., 2014) version 0.36 in ILLUMINACLIP MAXINFO mode. FastQC (Andrews, 2015) version 0.11.8 was used for quality control before and after adapter clipping, showing the successful removal of adapter content and consistently high read-quality of all samples. Cumulative quality reports for all datasets were generated with MultiQC (Ewels et al., 2016) version 1.6. The reference human genome build used for the genome alignment was GRCh38.p12 (Schneider et al. (2017)) and the reference bovine genome build was UMD.3.1.1 (Elsik et al. (2016)). DNA sequences were downloaded from Ensembl Genome Browser (Zerbino et al., 2018). Read alignment to the reference genomes was performed using Bowtie 2 (Langmead and Salzberg (2012)) version

2.3.4.1, with the local alignment option enabled and slightly relaxed constraints for the inter-mate distances of concordant read pairs, permitting fragment lengths between 100 and 600 bp. All other options were left at their default values.

2.6. Composition analysis

The analysis of the genomic alignments resulting from the mapping of the human sequencing data included chromosomes 1–22 plus the X chromosome, omitting chromosome Y to reduce the effect of gender imbalance in the sample sets. Likewise, only the bovine chromosomes 1–29 plus the bovine X chromosome were included in the analysis of the bovine data. The software for the analyses was implemented in Java™, using the Broad Institute Picard library (Picard, 2009), which facilitated access to the SAM/BAM (Li et al., 2009) formatted alignment files. From the alignment of every sample i the coverage $cov_i(c, p)$, representing the number of mapped DNA fragments at every base-pair position p on chromosome c, was derived. Only concordantly mapped read pairs were considered, ensuring a high certainty of the region of origin. Coverage values were incremented from the beginning of the first read to the end of the second read of each concordantly mapped pair, where a potential gap between both reads was filled and double counting of potential overlaps was averted. Using the repeat annotations from the UCSC Repeat Masker database (Casper et al., 2018), the coverage values of each sample were subsequently assigned to a selected set of repeat-families or non-repetitive regions. If more than one repeat-family was annotated at a distinct particular position, the coverage was distributed equally over the involved repeat families. A composition vector was computed from these assigned coverage values for the respective sample, where each element of the vector corresponds to a relative coverage of a repeat family $cov_{Ri}(Rfam)$ or a relative coverage of non-repetitive regions.

The relative coverage of a repeat family for sample i is given by

$$cov_{Ri}(Rfam) = 100\% \bullet \frac{\sum_{\forall\, c,\, p \in P_{Rfam}} \frac{cov_i(c,p)}{n_{Rfam}(c,p)}}{\sum_{\forall\, c,p} cov_i(c,\, p)},$$

Where P_{Rfam} is the set of all positions where repeat family $Rfam$ is annotated, $n_{Rfam}(c, p)$ is the number of annotated repeat families at position p on chromosome c and $cov_i(c, p)$ is the number of mapped DNA fragments at position p on chromosome. In analyses where only a selected subset of repeat families was regarded, the coverage at the remaining annotated repeat families was pooled under “other repeats”. The full spectrum of annotated repeat families can be found in Supplemental **Table S1**.

Statistical analysis of the compositions was performed in R (R Development Core Team, 2011) version 3.5.2 using the package `compositions` (van den Boogaart and Tolosana-Delgado, 2008; Van Der Boogaart and Tolosana-Delgado, 2006) for compositional data analysis. Mean and standard deviations were computed using the class of compositions in Euclidean geometry (`rplus`) (van den Boogaart and Tolosana-Delgado, 2008). Ternary diagrams were created using the Aitchison class of compositions (`acomp`) (van den Boogaart and Tolosana-Delgado, 2008; Aitchison, 1982). Multivariate analysis of variances (MANOVA) was performed using the isometric log-ratio transformation (ilr) (Egozcue et al., 2003) function of the package `compositions` and the R built in `manova` function.

In order to facilitate the comparison of the fractions in cfDNA-compositions with the corresponding fractions of the genome a representation value given by

$$rep(Rfam) = 100\% \bullet \frac{\bar{cov}_R(Rfam)}{abundance(Rfam)}$$

was introduced, where $\bar{cov}_R(Rfam)$ is the mean relative coverage of a repeat family across all samples and $abundance(Rfam)$ is the percentual genomic abundance of the respective repeat family. Thus, the representation value is the ratio of the relative fraction of a repeat family in the composition of the cfDNA to its genomic abundance, which is considered as an expectation value.

The representation of the data as relative coverage depth for sample i given by

$$depth_{Ri}(Rfam) = \frac{\frac{cov_{Ri}(Rfam)}{abundance(Rfam)}}{\frac{cov_{Ri}(Non\ repetitive)}{abundance(Non\ repetitive)}},$$

was used in order to establish a baseline for the cfDNA coverage depth of a repeat family, where the coverage depth of non-repetitive genomic regions serves as a reference point.

2.7. Coverage analysis

In order to obtain invariance from differing numbers of read pairs and alignment rates, average-count normalization was applied to the fragment count data. Each count value was divided through the average coverage of the respective sample given by $\bar{cov}_i = \frac{\sum_{\forall\, c,p} cov_i(c,p)}{\text{Genomesize}}$, where Genomesize is the size of the respective genome in base-pairs. Thus, normalized coverage values higher than 1.0 implies above average coverage. Line charts, illustrating the mean normalized coverage and its variance over a selected genomic region, were created using the JFreeChart plotting library (Gilbert, 2007). Coverage analysis for 146 full-length, intact L1 elements in the human genome was performed using the annotations from L1Base 2 (Penzkofer et al., 2017).

The mean normalized coverage of the sample set from healthy human individuals was scanned for covered regions in order to determine their length distribution. Noise from several sources, such as sequencing errors, alignment errors and inevitable DNA contaminations from ruptured blood cells, wet-lab equipment and sequencing, did not allow for an exact demarcation of the boundaries of the covered regions. Therefore, a coverage threshold for the discrimination of covered regions from noise was established. Every continuous region exceeding this threshold was interpreted as covered. The analysis was performed repeatedly with successively decreasing threshold in order to determine a lower boundary above the noise level. Ultimately, a coverage threshold of 2.0 times the average coverage was chosen. For the evaluation of strongly covered regions, a coverage threshold of 100.0 was used. In addition, constraints for a minimum region length of 100 bp and a maximum region length of 2,000 bp were introduced. The minimum constraint was used to filter very short regions caused by weakly covered peaks, which surpassed the threshold only at their top. The maximum constraint was used to limits the focus of the analysis by omitting large regions occurring in very low numbers. A histogram illustrating the frequencies of region lengths was created using the JFreeChart plotting library (Gilbert, 2007).

2.8. Fragment-length analysis

The DNA fragment lengths were mostly distributed between 200 and 800 bp after sample preparation. After Nextera-Library preparation the fragment length distribution was mostly between 250 and 700 bp, where about 120 bp accounted for adapters. Therefore, the final fragment lengths were mostly distributed between 130 and 580 bp. Although the original fragment lengths were not preserved in the sequencing data due to the fragmentation step, differences in the length distribution of fragments mapped to distinct genomic elements could still be evaluated. Approximated fragment length distributions were plotted in the form of histograms using the JFreeChart plotting library (Gilbert, 2007).

Table 2

Composition of the cfDNA in the samples from healthy human individuals regarding all significantly over- and underrepresented repeat families. All other annotated repeat families where pooled in the "other repeats" fraction. The representation value displays the ratio between the mean value of a fraction in the cfDNA composition and its corresponding fraction on the genome. The relative coverage depth displays the ratio between the coverage depth of a fraction in the cfDNA composition and the coverage depth at non-repetitive genomic regions.

Repeat Family:	Genome:	cfDNA-Composition			Relative coverage depth	
		mean	stdev	Representation	mean	stdev
Alu:	10.05 %	15.79 %	± 3.90 %	**157 %**	425 %	± 220 %
L1:	16.98 %	27.83 %	± 3.08 %	**164 %**	417 %	± 133 %
SVA:	0.14 %	0.27 %	± 0.12 %	**199 %**	544 %	± 327 %
centr:	2.36 %	20.12 %	± 6.93 %	**854 %**	2400 %	± 1610 %
Simple repeat:	1.18 %	2.70 %	± 0.88 %	**226 %**	612 %	± 408 %
Satellite:	8.54×10^{-2} %	1.90 %	± 0.80 %	**2230 %**	6030 %	± 3700 %
telo:	8.81×10^{-3} %	4.39×10^{-4} %	± $3.33 \cdot 10^{-4}$ %	**4.97 %**	11.0 %	± 5.5 %
acro:	1.35×10^{-3} %	2.99×10^{-3} %	± 2.03×10^{-3} %	**221 %**	552 %	± 498 %
Other Repeats:	19.09 %	9.88 %	± 2.92 %	**51.7 %**	121 %	± 4.3 %
Non repetitive:	50.10 %	21.53 %	± 6.65 %	**43.0 %**	100 %	–

3. Results

3.1. Composition analysis

For the cfDNA composition of healthy human individuals (see Table 2), we observed elevated fractions for the L1, Alu and SVA (SINE (short interspersed nuclear element)/VNTR(variable number tandem repeats)/Alu) repeat families of RTEs, which depend on the L1 transposition machinery (Dewannieux et al., 2003; Raiz et al., 2012; Faulkner and Billon, 2018). The overrepresentation of these elements ranged from 150 % to 200 % of their respective fraction in the composition of the genome. A second group of elevated elements was formed by different types of satellite DNA and simple repeats. While constituting comparatively small fractions of the genome, they were present at particularly high levels, with Satellite repeats accounting for more than 2,000 % of their genomic abundance. Centromeric alpha satellites (centr) accounted for more than 800 % and macro-satellites (acro) and simple repeats accounted for 200 % of their respective fraction on the genome. Telomeric satellite repeats (telo) were underrepresented with less than 5% of the expected value. Non-repetitive DNA and all other annotated repeat families, which were pooled under "other repeats", comprised between 40 and 70 % of their respective fraction of the genome. An overview on the cfDNA composition of healthy human individuals regarding the full spectrum of annotated repeat families can be found in the Supplemental **Table S1** A comparison between the cfDNA compositions of female and male healthy individuals (see Supplemental **Table S2**) yielded no significant difference with a p-value of 0.7389. The evaluation of a potential gender bias was not performed for the bovine sample set, since only samples from female individuals were available to us.

Similar overrepresentation of RTE-containing repeat families and satellite DNA was also observed for bovine cfDNA (see Table 4), where L1 and Core-RTE repeats featured representation values of about 130 %, BovB-RTE repeats of more than 200 % and centromeric alpha satellite repeats (centr) of more than 1,300 %.

The comparison of the blood plasma cfDNA compositions of healthy human individuals and sepsis patients, focusing on the more abundant repeat families L1, Alu and centr, showed a distinctive shift to RTE-containing repeat families for sepsis patients (see Table 3). An illustration of the distribution of the individual data points in the compositional multidimensional space is shown in Fig. 1. The shift in composition was particularly pronounced when comparing the contents of Alu and centr repeats in the ternary diagrams. Carrying out a MANOVA resulted in a p-value of $1.003 \bullet 10^{-11}$, indicating a significant difference in the distribution of compositions between both groups (i.e. healthy individuals and sepsis patients).

The small number of cancer cell culture samples, with two samples each of three different cancer types, did not allow a significant statement about differences in their cfDNA composition (see Table 3) to be made. Apparently, the VCaP cell lines featured elevated levels of Alu repeat DNA, similar to the sepsis samples, while in both BeWo cell lines only L1 repeat DNA seemed to be elevated in comparison to the healthy state. The fraction of centr repeat DNA was similarly strongly overrepresented in the cancer cell lines as in the human plasma.

3.2. Coverage analysis

The threshold-based evaluation of the average coverage of the samples from healthy human individuals resulted in 622,101 regions being classified as consistently covered, which corresponded to approximately 8.9 % of the genome. The coverage at these regions accounted for about 56.6 % of the total mapped cfDNA, where the remaining quota was either attributed to noise or inconsistently covered regions (i.e. with subthreshold average coverage). The distribution of cfDNA coverage over the respective genome turned out to be highly non-uniform. Strong coverage patterns were often observed centred over particular repeat-regions and frequently spanned areas which were significantly longer than 200 bp (see Fig. 2**).**

Although the length of the covered regions ranged up to 140 kbp, the majority of the covered regions was short and regions longer than 2000 bp were very rare. Fig. 3**a** shows a histogram of region lengths from 100 to 2000 bp, where regions from 160 to 200 bp turned out to be the most common. Scanning the coverage for strongly covered regions using a threshold of 100 times the average coverage resulted in 1061 regions with lengths up to 6000 bp. While covering approximately 0.016 % of the genome, about 7.5 % of the total cfDNA coverage was mapped to these regions. 716 regions were mapped to centromeric repeats (centr) and 201 were mapped to simple repeats. Other regions were found at non-repetitive areas (50), satellite repeats (49), L1 repeats (21), Alu repeats (8) and TcMar-Tigger repeats (1). The remaining 17 regions were mapped to areas where more than one repeat family was annotated.

The coverage plots of the 146 full-length intact L1 elements from L1Base 2 (Penzkofer et al., 2017) consistently showed a similar coverage pattern, where a large area of 5–6 kbp is covered with coverage values between 2 and 20 times of the average coverage. Supplemental **Fig. S3** illustrates a selection of eight of these elements. The coverage plots of the L1 repeats of the 12 strongly covered L1 regions resulted in a more inhomogeneous picture, where strong coverage was in some cases only found at a subsection of the repeat. Supplemental **Fig. S4** shows four examples.

Table 3
Comparison of the cfDNA compositions of the human sample sets regarding selected repeat families in relation to their genomic abundances. All other annotated repeat families where pooled in the “other repeats” fraction. The representation value displays the ratio between the mean value of a fraction in the cfDNA composition and its corresponding fraction on the genome. The relative coverage depth displays the ratio between the coverage depth of a fraction in the cfDNA composition and the coverage depth at non-repetitive genomic regions.

Repeat Family:	Genome:	Category:	cfDNA-Composition			Relative coverage depth	
			mean	stdev	Representation	mean	stdev
Alu:	10.05 %	Healthy	15.79 %	± 3.90 %	**157 %**	425 %	± 220 %
		Sepsis	20.18 %	± 5.49 %	**201 %**	559 %	± 245 %
		VCaP	19.35 %	± 6.29 %	**192 %**	428 %	± 218 %
		BeWo1425	13.32 %	± 2.64 %	**133 %**	259 %	± 81.4 %
		BeWo1426	13.00 %	± 2.07 %	**129 %**	252 %	± 77.0 %
L1:	16.98 %	Healthy	27.83 %	± 3.08 %	**164 %**	417 %	± 133 %
		Sepsis	29.19 %	± 3.61 %	**172 %**	468 %	± 133 %
		VCAP	26.66 %	± 1.53 %	**157 %**	336 %	± 49.0 %
		BeWo1425	30.19 %	± 0.00 %	**178 %**	343 %	± 41.1 %
		BeWo1426	29.63 %	± 1.17 %	**175 %**	337 %	± 63.7 %
centr:	2.36 %	Healthy	20.12 %	± 6.93 %	**854 %**	2400 %	± 1610 %
		Sepsis	16.47 %	± 4.61 %	**699 %**	1960 %	± 849 %
		VCaP	11.60 %	± 0.68 %	**492 %**	1070 %	± 276 %
		BeWo1425	12.45 %	± 0.38 %	**528 %**	1020 %	± 153 %
		BeWo1426	13.16 %	± 0.82 %	**558 %**	1080 %	± 228 %
Other Repeats:	20.51 %	Healthy	14.73 %	± 2.35 %	**71.8 %**	177 %	± 37.7 %
		Sepsis	14.47 %	± 2.59 %	**70.5 %**	185 %	± 30.1 %
		VCaP	18.63 %	± 0.62 %	**90.8 %**	195 %	± 33.1 %
		BeWo1425	17.93 %	± 0.10 %	**87.4 %**	169 %	± 21.1 %
		BeWo1426	17.92 %	± 0.12 %	**87.4 %**	168 %	± 24.1 %
Non repetitive:	50.10 %	Healthy	21.53 %	± 6.65 %	**43.0 %**	100 %	–
		Sepsis	19.69 %	± 5.07 %	**39.3 %**	100 %	–
		VCaP	23.76 %	± 4.81 %	**47.4 %**	100 %	–
		BeWo1425	26.12 %	± 3.12 %	**52.1 %**	100 %	–
		BeWo1426	26.29 %	± 3.94 %	**52.5 %**	100 %	–

3.3. Fragment-length analysis

Using the mapped sequencing data, fragments from 100 to 600 bp could be reconstructed from concordantly mapped read pairs. The length distribution of more than 185 million mapped fragments from the sequencing data of the sample set from healthy human individuals is shown in Fig. 3**b**. The peak of the distribution was between 150 and 160 bp from where it decreased smoothly and in an exponential fashion. The fragment length distribution of full-length, intact L1 elements (Fig. 4a) and strongly covered L1 elements (Fig. 4b) was more or less identical with this distribution, while for centromeric satellite repeats (Fig. 4c) the distribution showed a trend towards longer fragments. The length distribution of fragments mapped to simple repeat regions (Fig. 4d) was significantly broader, showing a larger fraction of longer fragments.

4. Discussion

L1 elements belong to the non-long terminal repeat (non-LTR) class of retrotransposons (Faulkner and Billon, 2018) and self-amplify via a mechanism called “target-primed reverse transcription” (Luan et al., 1993). While L1 retrotransposons are able to self-amplify autonomously, the amplification of Alu und SVA is dependent on the L1 machinery of retrotransposition, where SVA amplification (Raiz et al., 2012) requires both L1 encoded proteins ORF1p and ORF2p, while Alu amplification (Dewannieux et al., 2003) requires ORF2p only. Their overrepresentation in the composition of human blood cfDNA, as shown in Table 2, indicates an involvement of the ORF2p endonuclease-reverse transcriptase (Feng et al., 1996; Mathias et al., 1991) in the generation of cfDNA molecules. Additionally, the strong fraction of RTEs in the cancer cell line supernatants, as shown in Table 3, indicates that these elements may be actively released. Active release of newly synthesized Alu DNA was already described by Stroun *et al.* 2001 (Stroun et al., 2001b) and the release of microvesicles containing L1 and Alu DNA by tumor cells was shown by Balaj et al. 2011 (Balaj et al., 2011). The strong presence of L1 and BovB repeats in the bovine cfDNA, as shown in Table 4, provides additional evidence for the involvement of reverse transcription in the active release of RTE DNA and suggests that the underlying mechanisms are not limited to human cfDNA, as BovB is a LINE-like element that also codes for a reverse transcriptase (Szemraj et al., 1995; Kordiš and Gubenšek, 1999).

One of the major questions raised by these results is how these DNA fragments originate before they are excreted in the form of microvesicles or exosomes. In the “target-primed reverse transcription”, as well as in the “snap-velcro” model of L1 retrotransposition proposed by Viollet et al. 2014 (Viollet et al., 2014), reverse transcription is only initiated after nicking of the target DNA during the insertion process. Hence, free DNA fragments could only originate through abortive insertion events. This cannot account for the amount of secreted RTE DNA, since those events do not occur in sufficiently high frequencies (Faulkner and Billon, 2018). A possible answer to this question could be provided by the work of Schwertz et al. 2018 (Schwertz et al., 2018), where it was shown that, in enucleated human platelets, the L1 ORF1 and ORF2 proteins are not only continuously expressed, but also associate with messenger RNA, where the ORF2p generates RNA-DNA hybrids through reverse transcription. The authors also pointed out that this mechanism controls protein expression on transcription level. The increased release of RTE DNA observed for sepsis patients, as shown in Table 3, supports the idea of a connection between gene regulation and the release of these elements. The question of if RNA-DNA hybrids could be excreted from the cell via exosomes, as described by Takahashi et al. 2017 (Takahashi et al., 2017), or are somehow processed to double-stranded DNA before being released into the blood stream, remains yet to be answered. Considering that in Balaj et al. 2011 (Balaj et al., 2011) mainly single-stranded DNA was found in microvesicles after RNase treatment, while in Kahlert et al. 2014 (Kahlert et al., 2014) predominantly double-stranded DNA molecules were reported for tumor-derived exosomes, both scenarios seem to be possible. If RTE DNA or RNA-DNA hybrids in exosomes or microvesicles are associated with ORF1p/ORF2p and integration competent, they could provide an

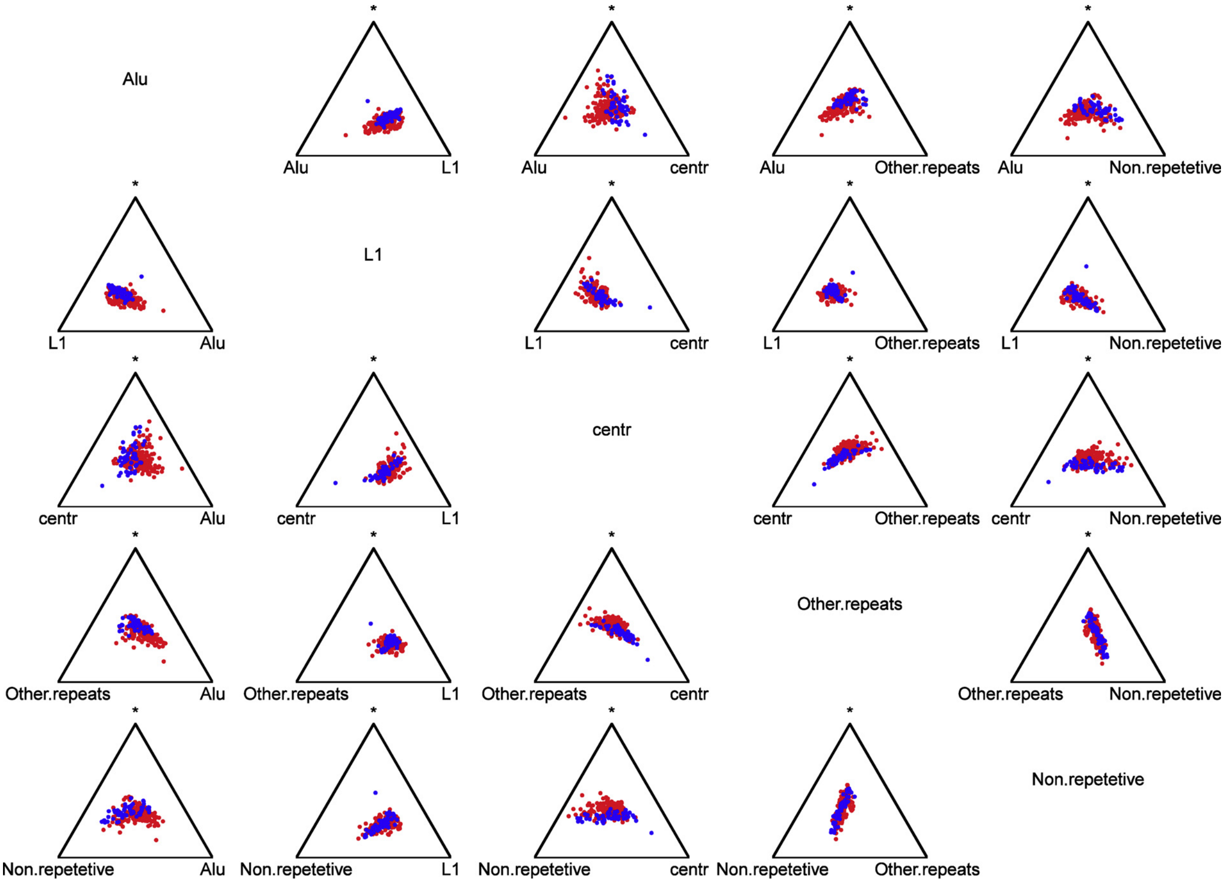


Fig. 1. Illustration of the compositions of human sepsis samples (red) and healthy samples (blue) in Aitchison geometry displayed as matrix of ternary diagrams. Each ternary diagram represents a reduction of the five-dimensional compositional space to three dimensions. Two individual components are plotted against a cumulation of the remaining components (indicated by *). Each point within a triangle corresponds to a valid three-component representation of a sample's composition, where the center of the triangle corresponds to a composition of equal parts and each corner of the triangle corresponds to 100 % of the respective component. Especially the Alu fraction was elevated in samples from sepsis patients in comparison to samples from healthy individuals. The details of the analysis are shown in Table 2 as well as in **Table S1** (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

explanation for the ability of cfDNA to integrate into the genome after uptake by cells as described in Mittra et al. 2015 (Mittra et al., 2015) and Gravina et al. 2016 (Gravina et al., 2016). Also, oncogenic transformation of cultured cells after exposure to isolated cfDNA (Thierry et al., 2016; Gravina et al., 2016; García-Olmo et al., 2010) could be possible under these circumstances.

Another group of overrepresented fractions in the cfDNA is comprised by different types of satellite DNA and simple repeat DNA. We initially tried to explain this observation with the association of those elements to constitutive heterochromatin (Csink and Henikoff, 1998; Plohl et al., 2008; Saksouk et al., 2015), which supposedly provides protection from digestion after cell death, but surprisingly, these

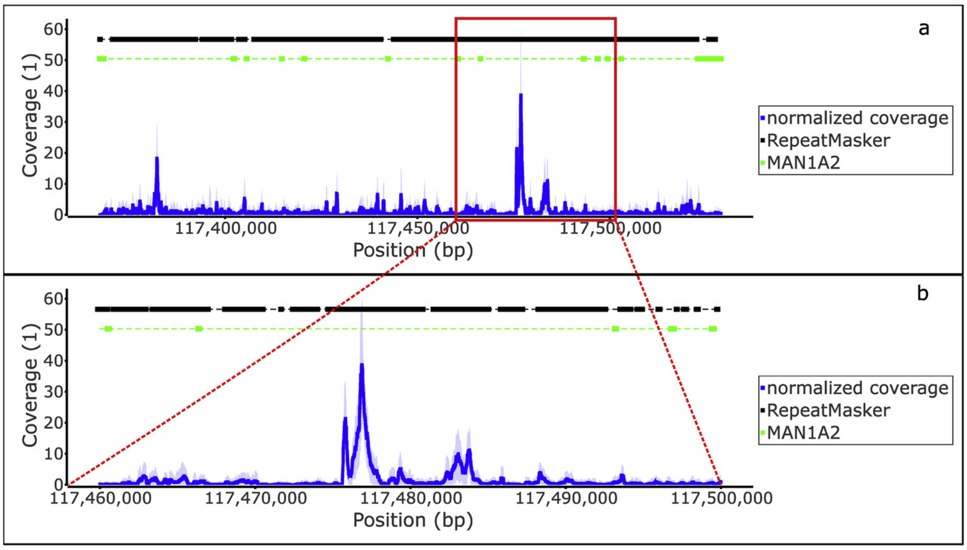


Fig. 2. Average normalized cfDNA coverage of the MAN1A2 gene on chromosome 1 of the samples from healthy human individuals (a). Black bars at the top indicate annotated repeat regions, while green bars indicate exon regions and dashed green lines indicate intron regions. The blue shaded area indicates the variance of the coverage. A closer look at the coverage peak on the intron within the red box is given by the lower section of the figure (b). The strong coverage peaks are positioned on repeats of the L1 repeat family (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

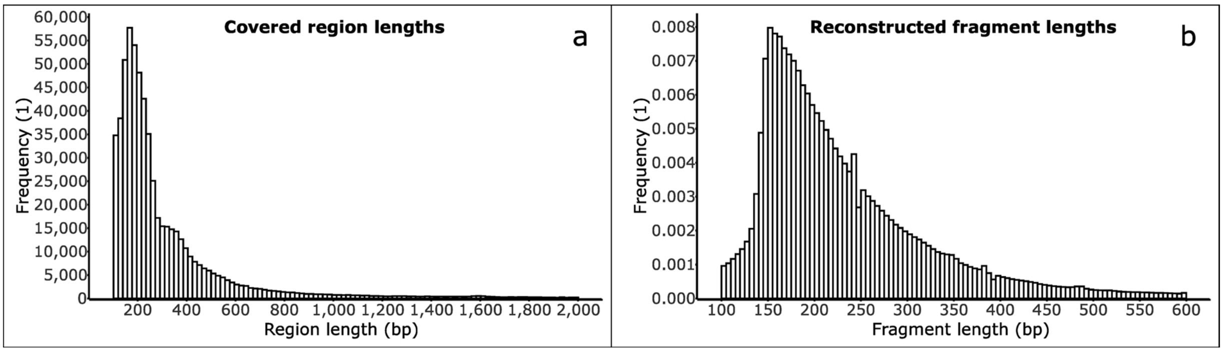


Fig. 3. Region and fragment lengths for the average coverage of the samples from healthy human individuals. Section (a) shows the Histogram of estimated lengths of regions covered by cfDNA. The ordinate shows absolute frequencies. Length estimation of the regions was performed with a threshold of 2.0 times the average coverage. Section (b) shows the fragment length distribution of all reconstructed fragments. The ordinate shows relative frequencies, where all bars sum to one.

elements were also strongly present in the supernatant of the cancer cell lines, indicating an active release instead. The especially high representation values of satellite and simple repeat DNA and the differences in their fragment length distribution suggest that a different mechanism is responsible for their presence when compared to RTE fragments. A potential source of nucleosolic or cytosolic repeat DNA fragments might be DNA replication. Centromeric satellite sequences tend to cause intermediate secondary structure formation on Okazaki fragments, stalling the replication fork (Li et al., 2018). These secondary structures are reportedly removed by the (h)DNA2 nuclease/helicase complex (Li et al., 2018; Bae et al., 1998; Pinto et al., 2016), resulting in DNA fragments which could be excreted via exosomes. Impediment of replication fork progression was also described for simple DNA repeat blocks (Krasilnikova et al., 1998), where the efficiency of the blockage was reported to be dependent on repeat length, orientation and simultaneous transcription.

The reason for the almost complete lack of telomeric satellite repeat DNA in the cfDNA composition (see Table 2) might be their mode of replication. Telomeres are tightly bound to constitutive heterochromatin (Saksouk et al., 2015) and telomeric repeats have been reported to be actively transcribed (Luke and Lingner, 2009; Maicher et al., 2012). This does not seem to promote the presence of telomeric satellite repeat DNA (telo) in the cfDNA population. Telomeres form a specialized heterochromatin structure (Saksouk et al., 2015) and their replication involves telomere-specific proteins (Gu et al., 2012; Martínez and Blasco, 2015). Telomeric Okazaki-fragments are synthesized in a distinct process, which differs from the conventional lagging-strand replication (Martínez and Blasco, 2015; Huang et al., 2012). DNA2 helicase/nuclease was reported to resolve replication fork stalling at G-quadruplex structures (Lin et al., 2013; Choe et al., 2002), whereas an involvement in Okazaki-fragment processing at telomeres apparently has not been reported thus far.

The observation that RTE and satellite DNA are consistently overrepresented in both human and bovine cfDNA (with no significant gender dependency observed in human samples, see **Table S2**) indicates that these results are most likely not related to effects of the DNA amplification and sequencing procedure. Additionally, the impact of eventual uneven amplification/sequencing of individual sequence elements should have been largely compensated by the marginalization of the coverage data over the large amounts of genomic loci aggregated

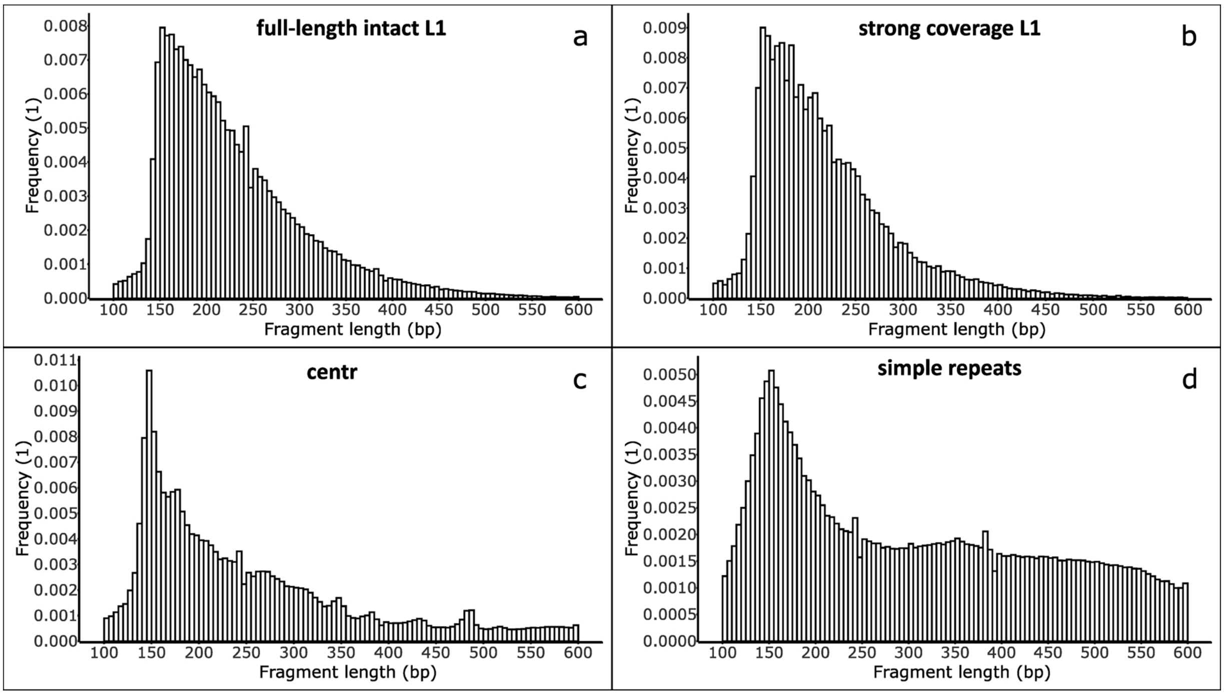


Fig. 4. Comparison of the fragment length distributions of cfDNA fragments mapped to full-length intact L1 elements (a), L1 elements featuring strong coverage (b), centr repeats (c) and simple repeat areas (d). The ordinates show relative frequencies, where all bars sum to one.

Table 4
Composition of the cfDNA in the bovine sample set regarding selected repeat families in relation to their genomic abundances. All other annotated repeat families where pooled in the “other repeats” fraction. The representation value displays the ratio between the mean value of a fraction in the cfDNA composition and its corresponding fraction on the genome. The relative coverage depth displays the ratio between the coverage depth of a fraction in the cfDNA composition and the coverage depth at non-repetitive genomic regions.

Repeat Family:	Genome:	cfDNA-Composition			Relative coverage depth	
		mean	stdev	Representation	mean	stdev
RTE-BovB:	12.21 %	26.35 %	± 9.42 %	**216 %**	438 %	± 296 %
L1:	12.80 %	17.20 %	± 2.17 %	**134 %**	243 %	± 78.4 %
centr:	0.27 %	3.65 %	± 1.92 %	**1370 %**	2660 %	± 1910 %
Core-RTE:	2.35 %	3.12 %	± 0.97 %	**133 %**	260 %	± 154 %
Other Repeats:	20.61 %	19.05 %	± 2.14 %	**92.4 %**	165 %	± 35.2 %
Non repetitive:	51.76 %	30.64 %	± 8.21 %	**59.2 %**	100 %	–

within a repeat family.

Our results suggest that a large fraction of cfDNA in serum/plasma is the result of active DNA release via exosomes into the blood stream. This is consistent with the observations described in the work of Fernando et al. 2017 (Fernando et al., 2017), where they stated that more than 90 % of cfDNA was localized in exosomes. The lack of a ladder-like pattern (Rumore and Steinman, 1990; Giacona et al., 1998) in the size distribution of covered regions and reconstructed DNA fragments shown in Fig. 3 could be explained by a majority of exosome-associated DNA fragments, with apoptosis derived cfDNA fragments, which would be wound around histones (Snyder et al., 2016), being largely outnumbered. Since most repeat families and nonrepetitive genomic elements together constituted about 40 % in the average cfDNA composition, we conclude that a large fraction of these DNA fragments is contained in exosomes as well. In our opinion, their representation at values around 40%–70%, when compared to their genomic abundance, suggests that no specialized mechanism is responsible for their occurrence. DNA fragments originating from apoptosis, DNA damage, DNA repair and hairpin formation during replication, which supposedly are also excreted from cells, may be a possible explanation for the presence of these genomic elements in the blood stream. Considering that a large fraction of the circulating DNA is being contained in exosomes and the significantly increased fraction of Alu repeat DNA in the cfDNA composition of blood plasma from sepsis patients, we would like to point out that exosome associated RTE DNA constitutes a promising target for the search for diagnostic biomarkers.

CRediT authorship contribution statement

Stefan Grabuschnig: Conceptualization, Software, Investigation, Writing - original draft, Writing - review & editing. **Jung Soh:** Data curation, Validation. **Petra Heidinger:** Methodology, Validation. **Thorsten Bachler:** Methodology, Validation. **Elisabeth Hirschböck:** Methodology, Validation. **Ingund Rosales Rodriguez:** Validation, Writing - review & editing, Resources. **Daniel Schwendenwein:** Validation, Writing - review & editing, Resources. **Christoph W. Sensen:** Conceptualization, Project administration, Supervision, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that there are no competing financial interests.

Acknowledgement

We want to thank Dr. Laura Villanova (TU Graz) for guidance in R and Dr. Maria Sensen, Dr. Jürgen Hartler and Maximilian Malek for critical reading of the manuscript. The Java™ source code for all analyses can be found on github (https://github.com/SG-ICBT/cnacomp). Special thanks goes to CNA Diagnostics Inc. who provided the sequencing data and agreed on publication of the datasets for healthy human individuals, bovina and cancer cell cultures. The genomic alignments created for these datasets are publicly available at the BioProject accession **PRJNA596372**. This paper is dedicated to the memory of Prof Dr. Robert Bertram Church, Calgary, AB, Canada, who passed away on Sep. 06, 2019.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jbiotec.2020.03.002.

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B 44, 139–160.

Andrews, S., 2015. FASTQC a quality control tool for high throughput sequence data. Babraham Inst.

Anker, P., Stroun, M., Maurice, P.A., 1975. Spontaneous release of DNA by human blood lymphocytes as shown in an in vitro system. Cancer Res. 35, 2375–2382.

Arends, M.J., Morris, R.G., Wyllie, A.H., 1990. Apoptosis. The role of the endonuclease. Am. J. Pathol. 136, 593–608.

Aucamp, J., Bronkhorst, A.J., Badenhorst, C.P.S., Pretorius, P.J., 2018. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biol. Rev. Camb. Philos. Soc. 93, 1649–1683.

Bae, S.H., et al., 1998. Dna2 of Saccharomyces cerevisiae possesses a single-stranded DNA- specific endonuclease activity that is able to act on double-stranded dna in the presence of ATP. J. Biol. Chem. 273, 26880–26890.

Balaj, L., et al., 2011. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nat. Commun. 2.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Breveglieri, G., D'Aversa, E., Finotti, A., Borgatti, M., 2019. Non-invasive prenatal testing using fetal DNA. Mol. Diagnosis Ther. 23, 291–299.

Bronkhorst, A.J., et al., 2016. Characterization of the cell-free DNA released by cultured cancer cells. Biochim. Biophys. Acta - Mol. Cell Res. 1863, 157–165.

Casper, J., et al., 2018. The UCSC genome browser database: 2018 update. Nucleic Acids Res. 46, D762–D769.

Chan, K.C.A., et al., 2013a. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin. Chem. 59, 211–224.

Chan, K.C.A., et al., 2013b. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc. Natl. Acad. Sci. 110, 18761–18768.

Choe, W., Budd, M., Imamura, O., Hoopes, L., Campbell, J.L., 2002. Dynamic localization of an okazaki fragment processing protein suggests a novel role in telomere replication. Mol. Cell. Biol. 22, 4202–4217.

Cocucci, E., Racchetti, G., Meldolesi, J., 2009. Shedding microvesicles: artefacts no more. Trends Cell Biol. 19, 43–51.

Csink, A.K., Henikoff, S., 1998. Something from nothing: the evolution and utility of satellite repeats. Trends Genet. 14, 200–204.

Danielson, K.M., Rubio, R., Abderazzaq, F., Das, S., Wang, Y.E., 2017. High throughput sequencing of extracellular RNA from human plasma. PLoS One 12.

Dewannieux, M., Esnault, C., Heidmann, T., 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

Elsik, C.G., et al., 2016. Bovine genome database: new tools for gleaning function from the Bos taurus genome. Nucleic Acids Res. 44, D834–9.

Ewels, P., Magnusson, M., Lundin, S., Kaller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048.

Faulkner, G.J., Billon, V., 2018. L1 retrotransposition in the soma: a field jumping ahead. Mob. DNA 9.

Feng, Q., Moran, J.V., Kazazian, H.H., Boeke, J.D., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905–916.

Fernando, M.R., Jiang, C., Krzyzanowski, G.D., Ryan, W.L., 2017. New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes. PLoS One 12.

Fleischhacker, M., Schmidt, B., 2007. Circulating nucleic acids (CNAs) and cancer-A survey. Biochim. Biophys. Acta Rev. Cancer 1775, 181–232.

Gahan, P.B., Stroun, M., 2010. The virtosome-a novel cytosolic informative entity and intercellular messenger. Cell Biochem. Funct. 28, 529–538.

García-Olmo, D.C., et al., 2010. Cell-free nucleic acids circulating in the plasma of colorectal cancer patients induce the oncogenic transformation of susceptible cultured cells. Cancer Res. 70, 560–567.

Giacona, M.B., et al., 1998. Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls. Pancreas 17, 89–97.

Gilbert, D., 2007. The JFreeChart Class Library. 42. .

Gravina, S., Sedivy, J.M., Vijg, J., 2016. The dark side of circulating nucleic acids. Aging Cell 15, 398–399.

Gu, P., et al., 2012. CTC1 deletion results in defective telomere replication, leading to catastrophic telomere loss and stem cell exhaustion. EMBO J. 31, 2309–2321.

Holdenrieder, S., et al., 2001. Nucleosomes in serum of patients with benign and malignant diseases. Int. J. Cancer 95, 114–120.

Holdenrieder, S., et al., 2005. Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. Clin. Chem. 51, 1544–1546.

Huang, C., Dai, X., Chai, W., 2012. Human Stn1 protects telomere integrity by promoting efficient lagging-strand synthesis at telomeres and mediating C-strand fill-in. Cell Res. 22, 1681–1695.

Jiang, P., Lo, Y.M.D., 2016. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. Trends Genet. 32, 360–371.

Kahlert, C., et al., 2014. Identification of doublestranded genomic dna spanning all chromosomes with mutated KRAS and P53 DNA in the serum exosomes of patients with pancreatic cancer. J. Biol. Chem. 289, 3869–3875.

Kalluri, R., LeBleu, V.S., 2016. Discovery of double-stranded genomic DNA in circulating exosomes. Cold Spring Harb. Symp. Quant. Biol. 81, 275–280.

Kordiš, D., Gubenšek, F., 1999. Molecular evolution of Bov-B LINEs in vertebrates. Gene 238, 171–178.

Korenchuk, S., et al., 2001. VCaP, a cell-based model system of human prostate cancer. In Vivo (Brooklyn). 15, 163–168.

Krasilnikova, M.M., M. Samadashwily, G., Krasilnikov, A.S., Mirkin, S.M., 1998. Transcription through a simple DNA repeat blocks replication elongation. EMBO J. 17, 5095–5102.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Lee, Y., El Andaloussi, S., Wood, M.J.A., 2012. Exosomes and microvesicles: extracellular vesicles for genetic information transfer and gene therapy. Hum. Mol. Genet. 21.

Li, H., et al., 2009. The sequence alignment/map format and SAM tools. Bioinformatics 25, 2078–2079.

Li, Z., et al., 2018. hDNA2 nuclease/helicase promotes centromeric DNA replication and genome stability. EMBO J. 37, e96729.

Lin, W., et al., 2013. Mammalian DNA2 helicase/nuclease cleaves G-quadruplex DNA and is required for telomere integrity. EMBO J. 32, 1425–1439.

Lo, Y.M.D., et al., 2007. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc. Natl. Acad. Sci. 104, 13116–13121.

Luan, D.D., Korman, M.H., Jakubczak, J.L., Eickbush, T.H., 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72, 595–605.

Luke, B., Lingner, J., 2009. TERRA: Telomeric repeat-containing RNA. EMBO J. 28, 2503–2510.

Maicher, A., Kastner, L., Dees, M., Luke, B., 2012. Deregulated telomere transcription causes replication-dependent telomere shortening and promotes cellular senescence. Nucleic Acids Res. 40, 6649–6659.

Majno, G., Joris, I., 1995. Apoptosis, oncosis, and necrosis. An overview of cell death. Am. J. Pathol. 146, 3–15.

Mandel, P., Metais, P., 1948. Les acides nucleiques du plasma sanguin chez l' homme [The nucleic acids in blood plasma in humans]. C. R. Seances Soc. Biol. Fil. 142, 241–243.

Martínez, P., Blasco, M.A., 2015. Replicating through telomeres: a means to an end. Trends Biochem. Sci. 40, 504–515.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Boeke, J.D., Gabriel, A., 1991. Reverse transcriptase encoded by a human transposable element. Science (80-.) 254, 1808–1810.

Mayers, J.R., Audhya, A., 2012. Vesicle formation within endosomes: an ESCRT marks the spot. Commun. Integr. Biol. 5, 50–56.

McCoubrey-Hoyer, A., Okarma, T.B., Holman, H.R., 1984. Partial purification and characterization of plasma DNA and its relation to disease activity in systemic lupus erythematosus. Am. J. Med. 77, 23–34.

Millan, C., Zenobi-wong, M., Akashi, M., 2014. Engineered cell manipulation for biomedical application. In: Akagi, T., Editors, M.M. (Eds.), Engineered Cell Manipulation for Biomedical Application, pp. 131–145. https://doi.org/10.1007/978-4-431-55139-3.

Mittra, I., et al., 2015. Circulating nucleic acids damage DNA of healthy cells by integrating into their genomes. J. Biosci. 40, 91–111.

Pattillo, R.A., Gey, G.O., 1968. The establishment of a cell line of human hormone-synthesizing trophoblastic cells in vitro. Cancer Res. 28, 1231–1236.

Penzkofer, T., et al., 2017. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 45, D68–D73.

Picard Package. Broad Institute. Picard Tools - By Broad Institute. Github (2009).

Pinto, C., Kasaciunaite, K., Seidel, R., Cejka, P., 2016. Human DNA2 possesses a cryptic DNA unwinding activity that functionally integrates with BLM or WRN helicases. Elife 5.

Plohl, M., Luchetti, A., Meštrović, N., Mantovani, B., 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409, 72–82.

R Development Core Team, 2011. R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computinghttps://doi.org/10.1007/978-3-540-74686-7.

Raiz, J., et al., 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res. 40, 1666–1683.

Rumore, P.M., Steinman, C.R., 1990. Endogenous circulating DNA in systemic lupus erythematosus. Occurrence as multimeric complexes bound to histone. J. Clin. Invest. 86, 69–74.

Saksouk, N., Simboeck, E., Déjardin, J., 2015. Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin 8.

Schneider, V.A., et al., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864.

Schwertz, H., et al., 2018. Endogenous LINE-1 (Long interspersed nuclear Element-1) reverse transcriptase activity in platelets controls translational events through RNA-DNA hybrids. Arterioscler. Thromb. Vasc. Biol. 38, 801–815.

Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M., Shendure, J., 2016. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 164, 57–68.

Stroun, M., Anker, P., Lyautey, J., Lederrey, C., Maurice, P.A., 1987. Isolation and characterization of DNA from the plasma of cancer patients. Eur. J. Cancer Clin. Oncol. 23, 707–712.

Stroun, M., et al., 1989. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology 46, 318–322.

Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A., Anker, P., 2001a. About the possible origin and mechanism of circulating DNA apoptosis and active DNA release. Clin. Chim. Acta 313, 139–142.

Stroun, M., Lyautey, J., Lederrey, C., Mulcahy, H.E., Anker, P., 2001b. Alu repeat sequences are present in increased proportions compared to a unique gene in Plasma/Serum DNA. Ann. N. Y. Acad. Sci. 945, 258–264.

Suraj, S., Dhar, C., Srivastava, S., 2016. Circulating nucleic acids: an analysis of their occurrence in malignancies. Biomed. Reports 6, 8–14.

Szemraj, J., Płucienniczak, G., Jaworski, J., Płucienniczak, A., 1995. Bovine Alu-like sequences mediate transposition of a new site-specific retroelement. Gene 152, 261–264.

Takahashi, A., et al., 2017. Exosomes maintain cellular homeostasis by excreting harmful DNA from cells. Nat. Commun. 8.

Thakur, B.K., et al., 2014. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell Res. 24, 766–769.

Thierry, A.R., El Messaoudi, S., Gahan, P.B., Anker, P., Stroun, M., 2016. Origins, structures, and functions of circulating DNA in oncology. Cancer Metastasis Rev. 35, 347–376.

Ullrich, E., et al., 2020. Evaluation of host-based molecular markers for the early detection of human Sepsis. J. Biotechnol. 310, 80–88.

Umu, S.U., et al., 2018. A comprehensive profile of circulating RNAs in human serum. RNA Biol. 15, 242–250.

van den Boogaart, K.G., Tolosana-Delgado, R., 2008. Compositions': a unified R package to analyze compositional data. Comput. Geosci. 34, 320–338.

Van Der Boogaart, K.G., Tolosana-Delgado, R., 2006. Compositional data analysis with 'R' and the package 'compositions'. Geol. Soc. London, Spec. Publ. 264, 119–127.

Viollet, S., Monot, C., Cristofari, G., 2014. L1 retrotransposition: the snap-velcro model and its consequences. Mob. Genet. Elements 4, e28907.

Wan, J.C.M., et al., 2017. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat. Rev. Cancer 17, 223–238.

Zerbino, D.R., et al., 2018. Ensembl 2018. Nucleic Acids Res. 46, D754–D761.

Table S1: Composition of the cfDNA in healthy human samples regarding all annotated repeat families in relation to the genomic abundances of the respective repeat families. The representation value displays the ratio between the mean value of a fraction in the cfDNA composition and its corresponding fraction on the genome. The relative coverage depth displays the ratio between the coverage depth of a fraction in the cfDNA composition and the coverage depth at non-repetitive genomic regions.

Repeat Family:	**Genome**	**cfDNA-Composition**			**Relative coverage depth**	
		mean	**stdev**	**Representation**	**mean**	**stdev**
Simple repeat:	1.18 %	2.67 %	± 0.87 %	**226 %**	6112 %	± 404 %
telo:	$8.81\cdot10^{-3}$ %	$4.29\cdot10^{-4}$ %	± $3.25\cdot10^{-4}$ %	**4.87 %**	10.8 %	± 5.42 %
L1:	16.98 %	27.83 %	± 3.05 %	**164 %**	417 %	± 131 %
hAT-Charlie:	1.50 %	0.77 %	± 0.23 %	**50.9 %**	119 %	± 2.24 %
MIR:	2.77 %	1.37 %	± 0.42 %	**49.6 %**	115 %	± 3.25 %
L2:	3.40 %	1.59 %	± 0.59 %	**46.8%**	107 %	± 9.75 %
CR1:	$3.92\cdot10^{-1}$ %	$1.89\cdot10^{-1}$ %	± $0.62\cdot10^{-1}$ %	**48.1 %**	111 %	± 4.87 %
ERVL-MaLR:	3.63 %	1.97 %	± 0.52 %	**54.3 %**	128 %	± 9.59 %
Alu:	10.05 %	15.79 %	± 3.86 %	**157 %**	425 %	± 217 %
hAT:	$3.74\cdot10^{-2}$ %	$1.57\cdot10^{-2}$ %	± $0.45\cdot10^{-2}$ %	**42.01 %**	98.4 %	± 4.21 %
ERVL:	1.85 %	0.98 %	± 0.31 %	**53.1 %**	123 %	± 4.94 %
LTR:	$2.25\cdot10^{-2}$ %	$1.15\cdot10^{-2}$ %	± $0.36\cdot10^{-2}$ %	**51.0 %**	118 %	± 7.02 %
ERV1:	2.69 %	1.43 %	± 0.34 %	**53.4 %**	128 %	± 18.3 %
hAT-Tip100:	2.78c	$1.25\cdot10^{-1}$ %	± $0.38\cdot10^{-1}$ %	**45.1 %**	105 %	± 2.27 %
Low_complexity:	$1.95\cdot10^{-1}$ %	$0.56\cdot10^{-1}$ %	$0.14\cdot10^{-1}$ %	**28.6 %**	67.9 %	± 5.46 %
TcMar-Tigger:	1.19 %	0.80 %	± 0.22 %	**67.1 %**	157 %	± 8.69 %
ERVL?:	$1.79\cdot10^{-2}$ %	$0.93\cdot10^{-2}$ %	± $0.33\cdot10^{-2}$ %	**51.8 %**	119 %	± 8.97 %
ERV1?:	$7.18\cdot10^{-3}$ %	$3.65\cdot10^{-3}$ %	± $1.27\cdot10^{-3}$ %	**50.9 %**	117 %	± 12.7 %
RTE-X:	$1.15\cdot10^{-1}$ %	$0.58\cdot10^{-1}$ %	± $0.19\cdot10^{-1}$ %	**50.5 %**	117 %	± 6.57 %
LTR?:	$3.83\cdot10^{-2}$ %	$1.86\cdot10^{-2}$ %	± $0.58\cdot10^{-2}$ %	**48.5 %**	113 %	± 6.65 %
ERVK:	$2.76\cdot10^{-1}$ %	$1.20\ \cdot10^{-1}$ %	± $0.29\ \cdot10^{-1}$ %	**43.7 %**	104 %	± 12.7 %
snRNA:	$1.11\cdot10^{-2}$ %	$0.69\cdot10^{-2}$ %	± $0.17\cdot10^{-2}$ %	**62.0 %**	147 %	± 13.7 %
MULE-MuDR:	$2.27\cdot10^{-2}$ %	$1.59\cdot10^{-2}$ %	± $0.48\cdot10^{-2}$ %	**70.1 %**	164 %	± 11.2 %
tRNA:	$1.23\cdot10^{-2}$ %	$0.62\cdot10^{-2}$ %	± $0.23\cdot10^{-2}$ %	**50.5 %**	115 %	± 10.8 %
DNA?:	$1.45\cdot10^{-2}$ %	$0.67\cdot10^{-2}$ %	± $0.25\cdot10^{-2}$ %	**46.4 %**	106 %	± 12.1 %
Gypsy:	$1.18\cdot10^{-1}$ %	$0.53\cdot10^{-1}$ %	± $0.18\cdot10^{-1}$ %	**44.9 %**	103 %	± 6.19 %
hAT-Blackjack:	$1.10\cdot10^{-1}$ %	$0.58\cdot10^{-1}$ %	± $0.18\cdot10^{-1}$ %	**52.7 %**	122 %	± 4.62 %
SVA:	0.14 %	0.27 %	± 0.12 %	**199 %**	544 %	± 323 %
Satellite:	$8.52\cdot10^{-2}$ %	1.90 %	± 0.73 %	**2230 %**	6040 %	± 3670 %
srpRNA:	$8.96\cdot10^{-3}$ %	$5.14\cdot10^{-3}$ %	± $1.06\cdot10^{-3}$ %	**57.3 %**	140 %	± 29.0%
hAT-Ac:	$1.26\cdot10^{-2}$ %	$0.56\cdot10^{-2}$ %	± $0.20\cdot10^{-2}$ %	**44.6 %**	102 %	± 9.83 %
rRNA:	$6.83\cdot10^{-3}$ %	$5.62\cdot10^{-3}$ %	± $2.07\cdot10^{-3}$ %	**82.4 %**	192 %	± 39.8 %
TcMar-Tc2:	$5.40\cdot10^{-2}$ %	$2.83\cdot10^{-2}$ %	± $0.81\cdot10^{-2}$ %	**52.4 %**	123 %	± 7.36 %
tRNA-Deu:	$2.07\cdot10^{-3}$ %	$0.99\cdot10^{-3}$ %	± $0.42\cdot10^{-3}$ %	**47.7 %**	107 %	± 18.7 %
RTE-BovB:	$4.26\cdot10^{-2}$ %	$2.46\cdot10^{-2}$ %	± $0.83\cdot10^{-2}$ %	**57.7 %**	133 %	± 8.74 %
5S-Deu-L2:	$8.88\cdot10^{-3}$ %	$4.44\cdot10^{-3}$ %	± $1.68\cdot10^{-3}$ %	**49.9 %**	113 %	± 13.8 %
PiggyBac:	$1.64\cdot10^{-2}$ %	$0.67\cdot10^{-2}$ %	± $0.21\cdot10^{-2}$ %	**41.1 %**	95.7 %	± 9.22 %
hAT?:	$1.00\cdot10^{-2}$ %	$0.44\cdot10^{-2}$ %	± $0.17\cdot10^{-2}$ %	**44.1 %**	100 %	± 14.4 %
Unknown:	$2.42\cdot10^{-2}$ %	$1.20\cdot10^{-2}$ %	± $0.43\cdot10^{-2}$ %	**49.4 %**	113 %	± 8.67 %
TcMar-Mariner:	$9.15\cdot10^{-2}$ %	$5.25\cdot10^{-2}$ %	± $1.63\cdot10^{-2}$ %	**57.4 %**	134 %	± 7.70 %
Helitron:	$1.21\cdot10^{-2}$ %	$0.66\cdot10^{-2}$ %	± $0.20\cdot10^{-2}$ %	**54.3 %**	127 %	± 10.3 %
Gypsy?:	$4.44\cdot10^{-2}$ %	$2.10\cdot10^{-2}$ %	± $0.75\cdot10^{-2}$ %	**47.3 %**	109 %	± 7.62 %
tRNA-RTE:	$2.28\cdot10^{-2}$ %	$1.38\cdot10^{-2}$ %	± $0.50\cdot10^{-2}$ %	**60.7 %**	139 %	± 11.6 %
hAT-Tip100?:	$9.00\cdot10^{-3}$ %	$4.42\cdot10^{-3}$ %	± $1.56\cdot10^{-3}$ %	**49.1 %**	113 %	± 8.76 %
DNA:	$9.73\cdot10^{-3}$ %	$4.51\cdot10^{-3}$ %	± $1.81\cdot10^{-3}$ %	**46.4 %**	104 %	± 15.3 %
Penelope:	$3.14\cdot10^{-3}$ %	$1.42\cdot10^{-3}$ %	± $0.52\cdot10^{-3}$ %	**45.1 %**	103 %	± 13.1 %
Dong-R4:	$3.76\cdot10^{-3}$ %	$2.07\cdot10^{-3}$ %	± $0.71\cdot10^{-3}$ %	**55.1 %**	127 %	± 15.0 %
scRNA:	$4.22\cdot10^{-3}$ %	$2.78\cdot10^{-3}$ %	± $0.53\cdot10^{-3}$ %	**65.9 %**	160 %	± 28.9 %
RNA:	$3.69\cdot10^{-3}$ %	$1.86\cdot10^{-3}$ %	± $0.76\cdot10^{-3}$ %	**50.4 %**	115 %	± 16.9 %
PiggyBac?:	$1.37\cdot10^{-3}$ %	$0.61\cdot10^{-3}$ %	± $0.26\cdot10^{-3}$ %	**44.2 %**	100 %	± 27.2 %
centr:	2.36 %	20.12 %	± 6.86 %	**854 %**	2400 %	± 1590 %
TcMar?:	$1.97\cdot10^{-3}$ %	$1.05\cdot10^{-3}$ %	± $0.41\cdot10^{-3}$ %	**53.7 %**	123 %	± 21.6 %
Helitron?:	$2.02\cdot10^{-3}$ %	$1.06\cdot10^{-3}$ %	± $0.33\cdot10^{-3}$ %	**52.5 %**	123 %	± 18.1 %
TcMar:	$9.52\cdot10^{-4}$ %	$4.53\cdot10^{-4}$ %	± $1.59\cdot10^{-4}$ %	**47.6 %**	113 %	± 27.2 %
hAT-Tag1:	$6.01\cdot10^{-4}$ %	$3.22\cdot10^{-4}$ %	± $1.64\cdot10^{-4}$ %	**53.6 %**	120 %	± 36.4 %
Merlin:	$5.51\cdot10^{-4}$ %	$4.14\cdot10^{-4}$ %	± $1.85\cdot10^{-4}$ %	**75.0 %**	171 %	± 40.3 %
TcMar-Pogo:	$1.31\cdot10^{-4}$ %	$0.65\cdot10^{-4}$ %	± $0.45\cdot10^{-4}$ %	**50.0 %**	110 %	± 59.3 %
SINE?:	$7.97\cdot10^{-5}$ %	$3.89\cdot10^{-5}$ %	± $3.08\cdot10^{-5}$ %	**48.8 %**	106 %	± 71.0 %
PIF-Harbinger:	$5.36\cdot10^{-5}$ %	$2.59\cdot10^{-5}$ %	± $1.97\cdot10^{-5}$ %	**48.4 %**	108 %	± 70.2 %
acro:	$1.35\cdot10^{-3}$ %	$2.99\cdot10^{-3}$ %	± $2.01\cdot10^{-3}$ %	**221 %**	552 %	± 493 %
Non repetitive:	50.10 %	21.53 %	± 6.59 %	**43.0 %**	100 %	-

Table S2: Comparison of the cfDNA compositions of healthy human females and males regarding all significantly over- and underrepresented repeat families in relation to their genomic abundances. All other annotated repeat families where pooled in the "other repeats" fraction. The representation value displays the ratio between the mean value of a fraction in the cfDNA composition and its corresponding fraction on the genome. The relative coverage depth displays the ratio between the coverage depth of a fraction in the cfDNA composition and the coverage depth at non-repetitive genomic regions.

Repeat Family:	Genome:	Category:	cfDNA-Composition			Relative coverage depth	
			mean	stdev	Representation	mean	stdev
Alu:	10.05 %	female	15.32 %	± 3.55 %	**152 %**	395 %	± 191 %
		male	16.35 %	± 4.29 %	**163 %**	461 %	± 249 %
L1:	16.98 %	female	28.71 %	± 2.13 %	**169 %**	421 %	± 134 %
		male	26.8 %	± 3.71 %	**158 %**	413 %	± 133 %
SVA:	0.14 %	female	0.27 %	± 0.12 %	**194 %**	501%	± 284 %
		male	0.28 %	± 0.12 %	**205 %**	594 %	± 370 %
centr:	2.36 %	female	19.14 %	± 5.82 %	**812%**	2170 %	± 1260 %
		male	21.26 %	± 8.04 %	**902 %**	2660 %	± 1930 %
Simple repeat:	1.18 %	female	2.56 %	± 0.62 %	**217 %**	561 %	± 305 %
		male	2.8 %	± 1.11 %	**237 %**	672 %	± 503 %
Satellite:	$8{,}54\cdot10^{-2}$ %	female	1.92 %	± 0.85 %	**2250 %**	5950 %	± 3890 %
		male	1.88 %	± 0.76 %	**2200 %**	6120 %	± 3540 %
telo:	$8.81\cdot10^{-3}$ %	female	$4.28\cdot10^{-4}$ %	± $3.2\cdot10^{-4}$ %	**4.84 %**	10.5 %	± 4.9 %
		male	$4.52\cdot10^{-4}$ %	± $3.54\cdot10^{-4}$ %	**5.11 %**	11.7%	± 6.19 %
acro:	$1.35\cdot10^{-3}$ %	female	$2.98\cdot10^{-3}$ %	± $1.69\cdot10^{-3}$ %	**221 %**	526 %	± 328 %
		male	$2.99\cdot10^{-3}$ %	± $2.4\cdot10^{-3}$ %	**222 %**	582 %	± 651 %
Other Repeats:	19.09 %	female	10.1 %	± 2.82 %	**52.9 %**	121 %	± 3.59 %
		male	9.62 %	± 3.08 %	**50.4 %**	121 %	± 5.14 %
Non repetitive:	50.10 %	female	21.97 %	± 6.39 %	**43.9 %**	100 %	-
		male	21 %	± 7.06 %	**41.9 %**	100 %	-

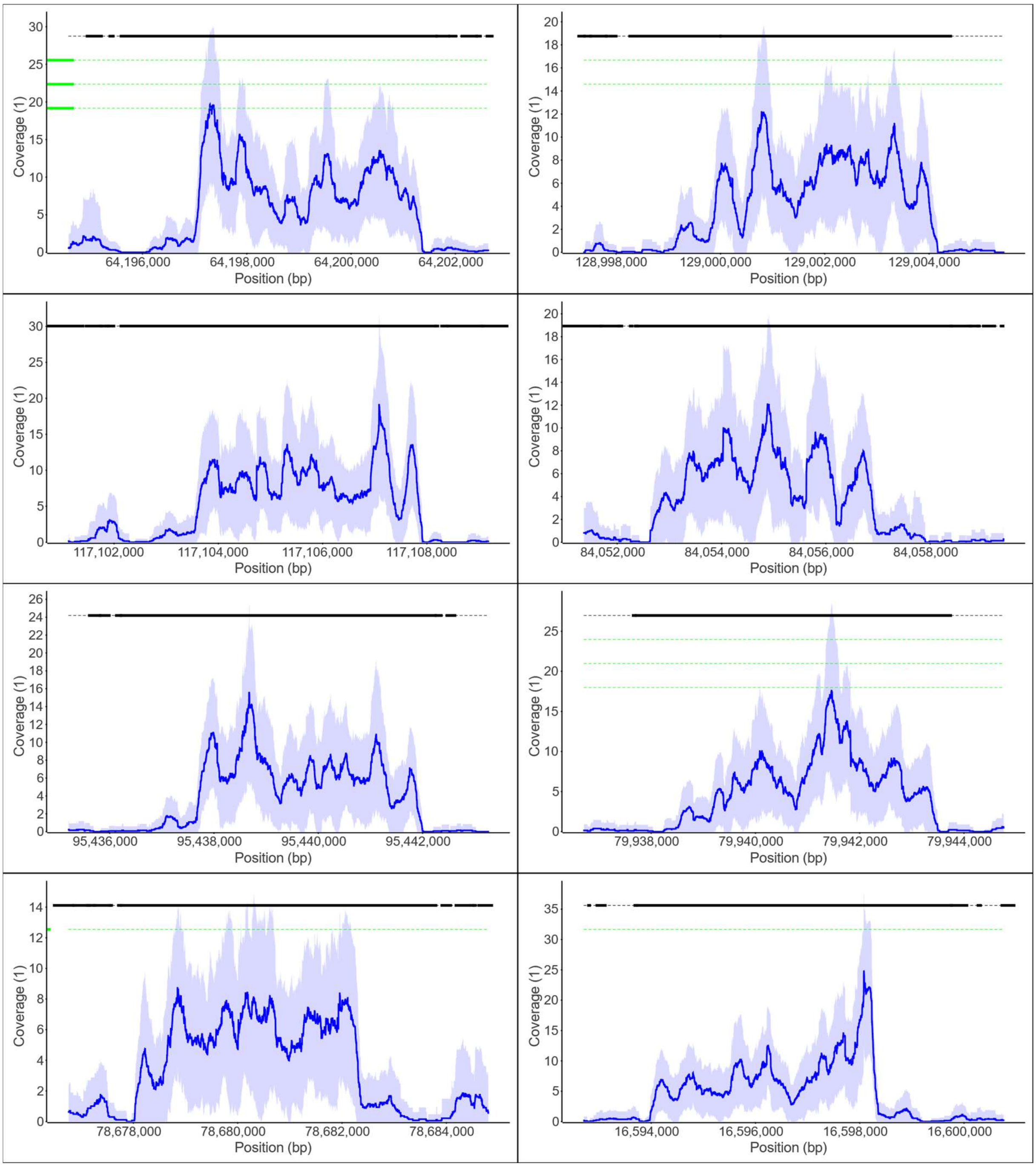


Figure S1: Coverage plots showing the mean normalized coverage of the 50 samples from healthy human individuals for 8 randomly selected full-length intact L1 elements from L1Base 2 (1).

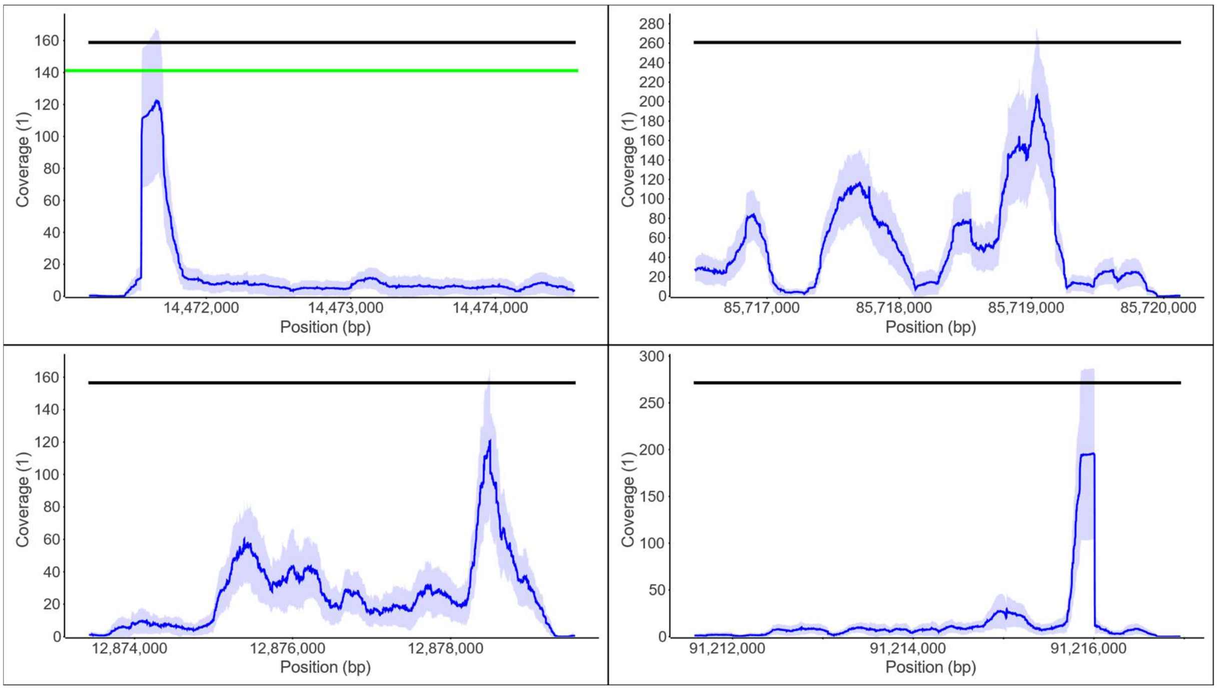


Figure S2: Coverage plots showing the mean normalized coverage of the 50 samples from healthy human individuals for 4 randomly selected L1 repeats featuring strong coverage.

Chapter 5

Identification and application of disease related cfDNA biomarkers

5.1 Prologue

The discovery and development of cfDNA biomarkers in blood serum or plasma of humans and animals for the diagnosis of disease conditions was the main scientific goal of our research group around my supervisor Prof. Dr. Christoph W. Sensen. The project was just in its beginnings when I started to work at the institute and was my personal entry point into the thematic on which I ultimately worked for the entirety of my PhD.

In an initial project phase we tried to identify disease specific cfDNA motifs in sequencing data from cattle afflicted with Johne's Disease, an infection of the small intestine in ruminants caused by *Mycobacterium avium ssp. paratuberculosis* (MAP) [81, 82, 83]. The contraction of this disease occurs mostly postnatally via faecal-oral transmission. It progresses very slowly and may also proceed without any clinical signs. The first symptoms usually arise at ages from 3 to 6 years in the form of chronic gastroenteritis and diarrhea. The disease onset then is followed by weight loss, emaciation and ultimately death. Currently, no treatment exists and disease control is difficult because it is poorly diagnosable before the manifestation of clinical signs [83]. As a consequence, Johne's Disease comprises a cause of substantial economic losses [84], especially in the Canadian and U.S. beef industry. The consumption of dairy products from MAP infected cattle is also suspected to be related to Crohn's Disease in humans [82, 83]. However, the difficulty to diagnose the disease posed a serious problem of finding a good control group of reliably uninfected cows, which was finally provided by the team of Prof. Dr. Lorenz Khol from the University of Veterinary Medicine, Vienna. Still, the certainty about positively diagnosed samples in early, preclinical infection stages was low and, in the end, we did not manage to achieve the desired diagnostic accuracy for a cfDNA biomarker-based blood test.

About this time we also began to work on another project concerning post-surgical sepsis in human patients. Fortunately, the knowledge about the ground truth behind these sample sets was much more reliable and we were also able to profit from our experiences gained during the initial project. In both projects Dr. Jung Soh and I were responsible for the Bioinformatics analyses of the cfDNA sequencing data and the identification of suitable candidate biomarker sequence motifs. At the time when we started, the task of identifying such biomarkers in whole-genome cfDNA sequencing data had not been done before. Thus, we had no prior knowledge of the problem dimensions and the challenges comprised by the nature and variability of our data as well as the extent of noise levels. Because of these uncertainties we agreed on each of us pursuing an individual approach independently instead of relying on a single strategy that might eventually be unfit for the problem. Therefore, we both successfully developed our individual software tools, producing candidate biomarker motifs. In the end, the signals which we were able to identify within the data from sepsis patients were much more pronounced than for Johne's Disease and looked very promising. A remaining problem was finding a method for the normalization of the corresponding Cq values from qPCR experiments to obtain invariance from sample properties such as differing water content. Since a concept like housekeeping genes, which are often used to standardize gene expression data, was unknown for cfDNA, we had to find an internal reference system which could be used for normalization. At first, we tried to identify regions with consistently stable coverage levels in all samples in order to use them as a reference, which did not work out very well. The breakthrough came via my side project regarding the composition of cfDNA (see Chapter 4), where it became apparent that different biological mechanisms seem to be responsible for the presence of different sequence elements within the cfDNA. Assuming that those may rise and fall independently, individual motifs cannot be used as a general reference. We therefore decided to evaluate the markers relatively to each other and tried to find pairings with high information content. Ultimately this idea led to a successful classification of samples from sepsis patients and healthy individuals.

5.2 Article

The article was accepted for publication in the Elsevier Journal of Biotechnology on January 27th 2020 and published on February 1st 2020.
DOI: 10.1016/j.jbiotec.2020.01.013

5.3 Author rights and permissions

Elsevier grants authors the rights to share and use their works for personal, scholarly or institutional purposes and explicitly permits inclusion of articles in theses and dissertations under the condition that a DOI Link to the original version on Science Direct is specified.
https://www.elsevier.com/about/policies/copyright

Journal of Biotechnology 310 (2020) 80–88

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Evaluation of host-based molecular markers for the early detection of human sepsis

Elisabeth Ullrich[a,1], Petra Heidinger[b,1], Jung Soh[c,1], Laura Villanova[d], Stefan Grabuschnig[d], Thorsten Bachler[b], Elisabeth Hirschböck[e], Sara Sánchez-Heredero[d], Barry Ford[f], Maria Sensen[g], Ingund Rosales Rodriguez[e], Daniel Schwendenwein[b], Peter Neumeister[h], Christoph J. Zurl[i], Robert Krause[j,l], Johannes Lorenz Khol[k], Christoph W. Sensen[d,e,l,*]

[a] *Institute of Hygiene, Microbiology and Environmental Medicine, Medical University of Graz, Neue Stiftingtalstraße 6, 8010, Graz, Austria*
[b] *acib GmbH, Petersgasse 14, 8010, Graz, Styria, Austria*
[c] *CNA Diagnostics Inc., Suite 300, 4838 Richard Road SW, Calgary, Alberta, T3E 6L1, Canada*
[d] *Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14/V, 8010, Graz, Styria, Austria*
[e] *CNA Diagnostics GmbH, Parkring 18, 8074, Grambach, Styria, Austria*
[f] *Defence Research and Development Canada, Suffield Research Centre, Suffield, P.O. box 4000 Stn Main, T1A 8K6, Medicine Hat, Alberta, Canada*
[g] *Hochstraße 12, 8076, Vasoldsberg, Styria, Austria*
[h] *Clinical Division for Hematology, Medical University of Graz, Auenbruggerplatz 38D, 8036 Graz, Styria, Austria*
[i] *Department of Pediatrics and Adolescent Medicine, Medical University of Graz, Auenbruggerplatz 34/II, 8036, Graz, Styria, Austria*
[j] *Section of Infectious Diseases and Tropical Medicine, Department of Internal Medicine, Medical University of Graz, Auenbruggerplatz 15, 8036, Graz, Styria, Austria*
[k] *Department for Farm Animals and Veterinary Public Health, University Clinic for Ruminants, University of Veterinary Medicine, Veterinärplatz 1, 1210, Vienna, Austria*
[l] *BioTechMed Graz, Mozartgasse 12/II, 8010, Graz, Styria, Austria*

ARTICLE INFO

Keywords:
Circulating nucleic acids
Host-based markers
Human sepsis diagnostics

ABSTRACT

We have identified 24 molecular markers, based on circulating nucleic acids (CNA) originating from the human genome, which in combination can be used in a quantitative real-time PCR (qPCR) assay to identify the presence of human sepsis, starting two to three days before the first clinical signs develop and including patients who meet the SEPSIS-3 criteria. The accuracy was more than 87 % inside of the same patient cohort for which the markers were developed and up to 81 % in blind studies of patient cohorts which were not included in the marker development. As our markers are host-based, they can be used to capture bacterial as well as fungal sepsis, unlike the current PCR-based tests, which require species-specific primer sets for each organism causing human sepsis. Our assay directly uses an aliquot of cell-free blood as the substrate for the PCR reaction, thus allowing to obtain the diagnostic results in three to four hours after the collection of the blood samples.

1. Introduction

CNA molecules, also referred to as cell-free DNA (cfDNA), are present in all mammalians. While the exact nature of their origin is still unclear (Aucamp et al., 2018; Thierry et al., 2016), the utilization of CNAs for the development of diagnostic approaches has already been subject to intense research, starting with cancer studies in the mid 1970′s (Leon et al., 1975) and leading to today's cancer detection and monitoring approaches, which have been termed “liquid biopsies” (Cheung et al., 2018; De Rubis et al., 2019; Poulet et al., 2019; Stewart et al., 2018; Stewart and Tsui, 2018). In addition to their use as cancer markers, CNAs have also been studied in the context of several other human diseases and conditions, including pregnancy complications (Bender et al., 2019; Gerson et al., 2019; Kumar and Singh, 2019), prenatal conditions (Renga, 2018; Van den Veyver, 2016); autoimmune diseases (Duvvuri and Lood, 2019; Truszewska et al., 2017); trauma (Ahmed et al., 2016; Gogenur et al., 2017; Jackson Chornenki et al., 2019; Thurairajah et al., 2018), stroke (Glebova et al., 2018; Vajpeyee et al., 2018), transplantation monitoring (Burnham et al., 2017; Gielis et al., 2019; Oellerich et al., 2016; Verhoeven et al., 2018), psychological stress (Trumpff et al., 2019) and even mental disorders (Jiang et al., 2018; Lindqvist et al., 2018).

Almost all high-resolution studies related to CNAs thus far have stayed on the DNA sequencing level and their resulting assays are

[*] Corresponding author at: Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14/V, 8010, Graz, Austria.
E-mail address: csensen@tugraz.at (C.W. Sensen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jbiotec.2020.01.013
Received 25 July 2019; Received in revised form 18 January 2020; Accepted 27 January 2020
Available online 01 February 2020

directly based on sequencing data from high-throughput DNA sequencing experiments. While this can be useful for slowly-developing conditions (*e.g.* many cancer types) as the analysis usually takes several days, DNA sequencing-based diagnostics is not feasible for highly dynamic diseases such as human sepsis, which fully develop within a few hours. In our study, we therefore focused on the development of quantitative real-time PCR (qPCR) assays, which would allow the completion of a diagnostic assay in less than 6 h in a diagnostic laboratory setting. It was our hypothesis that we should be able to transpose from high-throughput DNA sequencing experiments to qPCR assays in order to detect the development or presence of human sepsis using the cell-free fraction of blood (*i.e.* serum or plasma) as the substrate.

Detecting the onset of sepsis earlier, when compared to the currently available diagnostic methods, is one of the major challenges of health care providers, as human sepsis is a life-threatening condition, which can be caused by bacterial, fungal or viral infection. As reported on the sepsis fact sheet of the World Health Organization (World Health Organization, 2018), sepsis affects up to 30 million individuals annually worldwide, causing potentially 6 million deaths per year (Fleischmann et al., 2016). The direct cost of sepsis treatment in U.S. hospitals alone reached US$ 23.7 billion in 2013, which accounts for 6.2 % of the aggregate costs for all hospitalizations (Torio and Moore, 2016). Currently, blood cultures are still considered the “gold standard” for the detection of bacteremia/fungemia causing sepsis (Blevins and Bronze, 2010), although huge disadvantages and risks go along with them. This includes the time required for drawing up to three blood culture pairs and subsequent automatic culturing, low sensitivity in some entities like candidemia (Clancy and Nguyen, 2013) or infections caused by “culture-negative” microorganisms (Fenollar and Raoult, 2007; Gupta et al., 2016; Phua et al., 2013; Prost et al., 2013) (*e.g. Legionella sp.*, *Rickettsia sp.* or *Coxiella sp.*), as well as microorganisms with impaired growth due to species-specific traits or an already running anti-infective therapy. Contamination of blood cultures resulting in false positive results is also an important issue, as the most common organisms growing in blood cultures are coagulase-negative staphylococci, originating from the skin of patients (Abu-Saleh et al., 2018; Krause et al., 2003). One suggestion to overcome contamination would be to establish dedicated blood culture drawing teams available 24 h a day / 7 days a week, which would markedly increase personnel cost and is therefore rarely implemented in clinical routines (Peker et al., 2018).

Peptides, such as procalcitonin and presepsin, are the current “gold standard” molecular markers used in sepsis assays (Kondo et al., 2019). These tests can be used only once clinical signs are present (*i.e.* the SEPSIS-3 criteria have been met (Seckel, 2017)) and vary widely in their accuracy. To date, the U.S. Food and Drug Administration (FDA) has only approved the T2 Biosystems system (T2 Biosystems, 2020) for the direct detection of the presence of specific DNA molecules of pathogens in the blood of sepsis patients. Currently, this approach is severely limited by the fact that it is only able to detect a very limited number of bacterial and fungal species, which account for less than 50 % of all sepsis cases (Nguyen et al., 2019), thus missing all other pathogens causing sepsis. Like procalcitonin and presepsin, the T2 system is only used once clinical signs are present and cannot detect early stages of human sepsis. Complementing the current testing approaches with additional molecular techniques, which are capable of accurately identifying the onset of human sepsis earlier and in a more general way without missing the majority of patients, would therefore be a major paradigm shift in the diagnosis and treatment of human sepsis.

2. Materials and methods

2.1. Study cohorts

Both Medical University of Graz (MUG) study cohorts were age-restricted to age between 18 and 96 years. All of the individuals participating in the study were classified according to the SEPSIS-3 definition (Seckel, 2017). Clinical data used for the SEPSIS-3 assignment were derived from clinical databases and handwritten charts and extracted by a blinded, unrelated study physician not responsible for routine clinical assessment and creation of clinical data (*e.g.* assessment of the Glasgow Coma Score). The subsequent classification was performed by the blinded unrelated study physician and incorporated classification of preexisting organ dysfunction, as required for the SEPSIS 3 definition. None of the controls did match the criteria for SEPSIS-3 at any timepoint in the study.

At MUG, blood drawn for routine purpose (*e.g.* clinical chemistry, blood counts *etc.*) is stored for up to four days, enabling repetitive measurements in case of unreliable routine lab results. In case of positive blood cultures, we therefore had the opportunity to go back to retained blood samples drawn prior to index blood cultures in order to obtain the early timepoints required for our study. Blood cultures were ordered as soon as patients had signs and/or symptoms indicative for bacteremia/fungemia, as derived from current literature and described in local blood culture collection guidelines.

2.1.1. First study cohort for marker development collected at MUG

The initial results for the early detection of CNA-based sepsis markers were determined using 193 human blood serum/plasma samples, including 135 samples from 63 patients diagnosed with bacterial or fungal sepsis according to the SEPSIS-3 definition (*i.e.* infected) and 58 samples from 47 healthy individuals (*i.e.* controls: *e.g.* patients admitted to hospital for elective surgery with no infectious disease based on missing clinical signs/symptoms,lab results and blood sampling prior to surgery; in addition healthy volunteers, who were free of clinical symptoms or signs indicative of infection), as well as patients suffering from influenza and lymphoma, but not matching the SEPSIS-3 criteria from whom samples were collected as part of the NOBIS/NOBICS cohorts (Krause, 2020) at MUG. The collection timepoints of the sepsis case samples ranged from four days before the day on which the blood culture was ordered (designated as day 1) to three days afterwards. Samples from all available time points were included in the cohort, rather than aligning them to the time points presented in the second (test) cohort, to avoid artificial boosting of the test performance. Gram-negative or Gram-positive bacteria, as well as fungi, were identified as the sepsis-causing pathogens by routine measures including BACTEC blood cultures machines (Bactec FXTM, Becton Dickinson, Heidelberg, Germany) and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (Bruker Maldi Biotyper®, Bruker, Vienna, Austria).

Thus this cohort was carefully designed to facilitate the development of early DNA markers that are specific to sepsis due to bacteremia or fungemia, not just general disease conditions, as opposed to healthy conditions. Table 1 shows the details of the first cohort, which was obtained from MUG initially and used for marker development. Medical baseline characteristics of the first MUG cohort are shown in the Supplemental Tables S1a and S1b.

2.1.2. Second MUG study cohort for performance assessment

An additional set of 152 samples, consisting of 113 confirmed sepsis samples from 38 patients (according to the SEPSIS-3 standard, (Seckel, 2017)) and 39 control samples from 37 healthy individuals or patients suffering from a disease other than sepsis (for details see 2.1.1.), were used to evaluate the performance of our approach with a set of samples that was completely separate from the first MUG cohort and obtained from patients investigated in clinical routines. All of these samples were obtained from MUG after the marker development was finished and thus were not used during the marker development or the training of the classifiers. The timepoints of these sepsis samples ranged from four days before the day on which the blood culture was ordered to one day after the day of blood collection. Both Gram-negative and Gram-positive bacteria as well as fungi were identified as sepsis pathogens in these patients. Table 2 shows the details of this second MUG cohort.

Table 1
First MUG Patient Cohort.

Sample class	Pathogen (sepsis) or condition (comparator/control)		Patients	Samples	Samples taken at day(s) relative to the day when blood culture was drawn (=day 1)							
					–4	–3	–2	–1	1	2	3	4
Sepsis	Gram- negative bacteria	*Escherichia coli*	18	38	3	4	9	13	48	51	4	3
	Gram- positive bacteria	*Staphylococcus aureus*	19	43								
		Staphylococcus epidermidis	1	4								
	Fungi	*Candida spp.*	25	50								
Comparator/ Control	Healthy		38	38	Not applicable							
	Influenza		7	14								
	Lymphoma		2	6								

For both patient cohorts, the anonymized phenotypic description included detailed information on the disease state, which was used to classify all individual samples according to the SEPSIS-3 standard (Seckel, 2017). For a detailed summary see Supplemental Table S1a and S1b. All MUG samples were collected under Ethics approval No. 21-469 ex 09/10, all other samples were below the numbers, where a specific ethics protocol had to be established.

2.2. Plasma collection (Human Sepsis samples)

Blood was harvested into a syringe using a 21-gauge needle. The blood was transferred immediately to an 8 ml LH tube Lithium Heparin Sep, Greiner VACUETTE GBO Cat. No. 455,083. Following the blood collection, all tubes were gently inverted 5–10 times. The samples were centrifuged using a horizontal rotor (swing-out head) or a fixed angle rotor, for 10 − 15 min with a g-force of 1800 x g to 2000 x g. After the centrifugation step the separated plasma was aliquoted immediately and initially stored at +4 °C until plasma aliquots were frozen and stored long-term at −80 °C. A separate study of the stability of cfDNA in cattle serum samples showed that within a timeframe of 4 h at room temperature, the genomic maps changed by less than 10 %, thus allowing for some latitude in sample collection without degrading the DNA contained in the samples (data not shown).

2.3. DNA preparation for high-throughput DNA sequencing (Illumina)

Nucleic acids were extracted from plasma using the High Pure Viral Nucleic Acid Kit Roche Applied Sciences, 11858874001 according to the manufacturer's instructions. Subsequently, the extracted nucleic acids were amplified using the WGA4 GenomePlex Single Cell Whole Genome Amplification Kit Sigma-Aldrich, WGA4-500RXN according to the manufacturer's instructions. The amplified DNA was purified using the GenElute™ PCR Clean-Up Kit (Sigma-Aldrich, NA1020-1KT) according to the manufacturer's instructions and subsequently sent for high-throughput DNA sequencing to SEQ-IT GmbH, Kaiserslautern, Germany.

2.4. Illumina sequencing (performed by SEQ-IT GmbH, Kaiserslautern, Germany)

All amplicons were purified using the Agencourt AMPure®XP system on a BioMek NX workstation (Beckman Coulter) and quantified using a FluoStar Optima® (BMG Labtech) with the Quant-iT Picogreen® dsDNA reagent (Life technologies, Darmstadt, Germany). The DNA was diluted to a final concentration of 0.2 ng/μl. Library preparation for Illumina deep sequencing was done using the Nextera® XT DNA Sample Preparation and Index kit (Illumina, FC-131-1096), according to the manufacturer's instructions. The resulting libraries were normalized according to the manufacturer's instructions by SEQ-IT and pooled for subsequent sequencing on an Illumina NextSeq® 500 platform using the 2 × 150 cycle paired-end sequencing protocol.

2.5. Identification of informative genomic regions (Motifs)

Individual 150-bp long reads were extracted from the FASTQ sequence files produced by Seq-IT (see above). Adapter bases were

Table 2
Second MUG Patient Cohort.

Sample class	Pathogen (sepsis) or condition (comparator/control)		Patients	Samples	Samples taken at day(s) relative to the day when blood culture was drawn (=day 1)					
					–4	–3	–2	–1	1	2
Sepsis	Gram- negative bacteria	*Escherichia coli*	7	13	5	13	13	15	36	31
		Klebsiella oxytoca	1	2						
		Klebsiella variicola	1	2						
		Klebsiella pneumoniae	1	4						
		Stenotrophomonas maltophilia	1	5						
		Bacteroides thetaiotaomicron	1	2						
		Pseudomonas aeruginosa	1	2						
	Gram- positive bacteria	*Staphylococcus aureus*	8	24						
		Staphylococcus epidermidis	2	9						
		Enterococcus faecium	5	24						
		Streptococcus sp.	1	6						
		Streptococcus dysgalactiae	1	2						
		Streptococcus agalactiae	2	4						
		Streptococcus pneumoniae	2	4						
	Multiple bacteria		1	1						
	Fungi	*Candida spp.*	3	9						
Comparator/ Control	Healthy		33	33	Not applicable					
	Influenza		3	5						
	Lymphoma		1	1						

trimmed by using Cutadapt (Martin, 2011). The reads were then pair-merged and filtered using VSEARCH (Rognes et al., 2016) with minimum overlaps of 10 bp and a maximum of ambiguities within the overlapping region of three bp. Using software tools developed in-house, the read pairs were clustered into contigs of highly similar reads. For each sample, abundance counts were generated for the clusters, which were used to identify those clusters that had similar counts across all samples (universal motifs), those that were more than two-times higher in their abundance counts in sepsis patients (disease motifs) and also those that were more than two-times higher in their abundance counts in controls (control motifs). In a separate approach, the read pairs were mapped to the human genome assembly GRCh38.p12 (Schneider et al., 2017) and universal, disease-specific and control-specific motifs were identified through map comparisons using in-house software tools. The motifs used in our study and their main properties are listed in the Supplemental Table S2.

2.6. Generation of qPCR Cq values

qPCR was performed using a BioRad CFX96 Touch™ Real-Time PCR Detection System operating with the CFX Maestro™ 1.0 software (version 4.0.2325.0418, Bio-Rad Laboratories, USA) and using the Luna® Universal qPCR reaction kit (New England Biolabs, M3003E). The fluorescence signal of the PCR products was monitored continuously after each cycle, with the quantitation cycle (*Cq)* value determination mode set to "Regression Mode". An aliquot of 2 µl of a 1:40 diluted plasma sample was mixed with 10 µl of Luna® Master Mix, 0.5 µl of forward primer [10 µM], 0.5 µl of reverse primer [10 µM] (as suggested in the Luna® -qPCR instructions) and 7 µl of purified water (*Aqua bidest*, Fresenius). The PCR protocol consisted of an initial denaturation step at 95 °C for 2 min, followed by 40 cycles of 95 °C for 15 s and 45 s of annealing/extension at 60 °C. All data files derived from one study were combined and each *Cq* value generated was normalized by an inter-run calibration using a defined amount of synthethic DNA containing the target sequence which was included in every run, which was used as calibrator to normalize differences between individual PCR runs. In particular, the mean of all *Cq* values gained from each calibrator amplification was used for data normalization. Primer pairs with a cycle range between 15 and 41 were selected for the subsequent studies.

2.7. Identification of suitable primer pairs

Primer pairs were calculated using the Primer Quest Tool from Integrated DNA Technologies, USA using motif regions (see above), which resulted in PCR fragments of at least 120 bp in size. The primer size varied between 18 and 24 nucleotides. The primers were synthesized at Integrated DNA Technologies, USA and shipped to Graz, Austria, lyophilized. Evaluation of primer-pair specificity was performed by agarose-gel analysis and melting-curve analysis [60 °C–95 °C; increment of 0.5 °C for 5 s; plate read], whereas evaluation of primer-pair reliability was performed by investigation of qPCR *Cq* values. As PCR efficiency decreased when using serum samples, in this study a larger number of PCR cycles was allowed, when compared to standard analyses. Thus, qPCR *Cq* values in the 10–41 cycle range were considered acceptable, whereas values outside the 10–41 range were set to "Out of Range (OR)". Only primer pairs resulting in i) a single fragment of the correct size during the agarose-gel analysis ii) a single peak of the expected melting temperature for the anticipated PCR fragment and iii) 98 % of the measured *Cq* values in the 10–41 range were chosen for the subsequent analyses. A total of 24 primer pairs were selected. (referred to as "biomarkers" from here on). Detailed information about the primers is contained in the Supplemental Table S2. The DNA sequences described here were submitted for patent protection under ATTY. DOCKET No. 176395-010601/US/CON (Methods for Treating and Detecting Sepsis in Humans (Sensen et al., 2019)).

2.8. Generation of ΔCq values

In order to account for differences between samples, relative *Cq* values within the same sample were used for data analysis. Let $Cq_{A,i}$ denote the *Cq* value observed for biomarker *A* in sample *i*, and $Cq_{B,i}$ denote the *Cq* value observed for biomarker *B* in sample *i*; then, $\Delta Cq_{AB,i} = Cq_{A,i} - Cq_{B,i}$ indicates the relative aboundance of biomarker *A* compared to biomarker *B* observed for sample *i*. Using a script implemented in R (R-project, 2020). Δ*Cq* values were determined for all pairs of 24 biomarkers, namely $\binom{24}{2} = 276$ biomarkers pairs; thus, a set of 276 Δ*Cq* values was derived for each individual plasma/serum sample. Only linearly independent and linearly uncorrelated biomarker pairs with a Pearson Correlation Coefficient of 0.6 (or less) as cut-off (Karl and Erdmann, 1896) were retained resulting in 23 biomarker pairs. The 23 biomarker pairs were used as input variables in the subsequent classifier development step.

2.9. Classifier implementation

The ability of the 23 biomarker pairs (input variables) to discriminate between sepsis and control samples (output class) was investigated using various classifier algorithms. The implementation provided by the R-package caret (Classification And REgression Training, http://caret.r-forge.r-project.org) was used. A total of 92 different classifier algorithms were trained on the first MUG patient cohort and the five best-performing models were selected for testing on the second MUG cohort. The optimal classifier algorithms included: Support Vector Machines with Class Weights (Vapnik, 1998), Localized Linear Discriminant Analysis (Tutz and Binder, 2005), Self-Organizing Maps (Kohonen et al., 2001), Mixture Discriminant Analysis (Hastie and Tibshirani, 1996), and Neural Network (Ripley, 1996). Our study did not aim at selecting a final classification model; rather the information content of the CNA-based markers was evaluated by the classification performance.

2.10. Classifier performance assessment

For each study, the classifier performance measures of sensitivity, specificity, accuracy, positive predictive value and negative predictive value were determined. Fig. 1 shows the workflow for training and testing a classifier using two different datasets. For the characterization of each serum/plasma sample, there were four possible outcomes of the classification procedure:

- True positive: Sepsis sample correctly classified as sepsis,
- False positive: Control sample incorrectly classified as sepsis,
- True negative: Control sample correctly classified as control,
- False negative: Sepsis sample incorrectly classified as control.

First MUG cohort: To train a classifier on the MUG dataset, five-fold cross validation with five repetitions were performed. In this training scheme, the training dataset was randomly partitioned into five folds (Fold 1, Fold 2, Fold 3, Fold 4 and Fold 5), where each fold contained approximately the same number of samples. Subsequently, the classifier was trained on four folds and tested on the remaining fold, *i.e.* the one not used for training. This procedure of training on four folds, followed by testing on the remaining fold was done in five iterations, each time using a different set of four folds for training, such that each of the samples was tested for classification only once. The whole process was repeated for a total of five times, each time with a different partitioning, to reduce the potential influence of any particular random partitioning of training samples into five folds on the performance measures.

Additionally, when training a model different parametrizations were tested using a random search across a grid of parameters combinations. Parameters used and their values were set as per the default

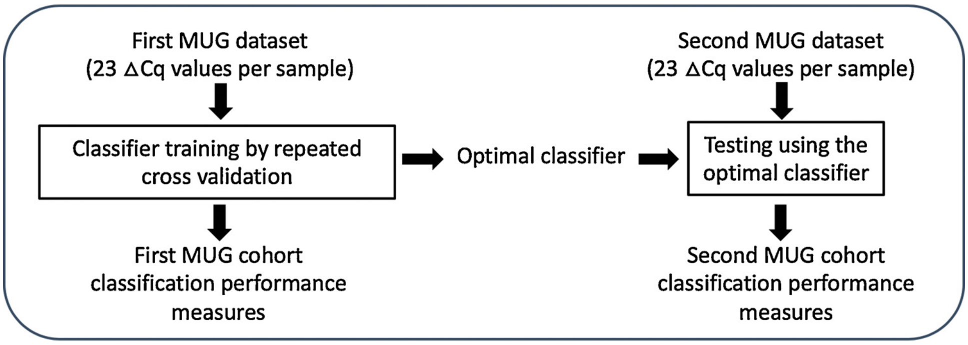

First MUG dataset
(23 ΔCq values per sample)
Classifier training by repeated cross validation
First MUG cohort classification performance measures
Optimal classifier
Second MUG dataset
(23 ΔCq values per sample)
Testing using the optimal classifier
Second MUG cohort classification performance measures

Fig. 1. Workflow for the performance evaluation of the classifiers used in our study.

values of the *trainControl* function from the R *caret* package, whereas the number of parameter combinations was set to five.

For each model, the performance measures from all repetitions and parametrizations were averaged to derive combined performance measures. Based on these combined performance measures, a subset of best-performing classifiers was selected for further testing and only the classifiers with the best-performing parametrization were selected (best classifiers).

Second MUG cohort: The best classifiers as trained on the first MUG dataset were used without modification for testing on the second MUG dataset.

3. Results

3.1. Mapping results

Mapping of the sequencing data to the reference genome resulted in high alignment rates to the human genome overall (between 95%–99%), indicating low contamination rates. The read coverage was distributed unevenly over the genome, with large regions (spanning up to several kilobase pairs) being very sparsely covered, or not covered at all. Protein-coding regions were mostly covered at a low level, while some intronic and repetitive regions featured peaks with consistently high coverage, as shown in Fig. 2. Differences in coverage levels between sepsis and control samples were observed for distinct peaks (also shown in Fig. 2). The read-count ratios of these peaks between patients and controls were the basis for the selection of marker regions. We have studied the nature of the informative CNA fragments, whose origin can be mainly linked to gene expression and described it in detail in a separate submission (Grabuschnig et al., 2019, submitted).

3.2. Comparison of genomic mapping vs. PCR results

In most cases, read count ratios were directly related to numbers of qPCR cycles, *i.e.* high read count numbers resulted in low *Cq* values and *vice versa*. Fig. 3A shows an example of this relationship between high-throughput DNA sequencing results and qPCR results. In a few instances though, the relationship between the sequence mapping results and the qPCR results were the opposite of the expectation, *i.e.* the qPCR cycle numbers were higher in instances of high read count numbers than in low read count number cases. Fig. 3B shows an example of this inverse relationship.

The motifs showing an inverse relationship between the high-throughput DNA sequencing results and the qPCR results were usually located in genomic areas, which are known to possess a high degree of secondary nucleic acid structures, such as ribosomal DNA genes.

3.3. Diagnostic classification performances on MUG patient cohorts

The average performance of all 92 classifiers, which were evaluated in this study, is shown in the Supplemental Table S3. From the 92 algorithms initially used, five exceeded 87 % overall balanced accuracy and resulted in balanced accuracy values that were stable across multiple timepoints. For these five algorithms, the performance data resulting from the best parametrization are shown in Table 3A and B. Table 4A and B shows the balanced accuracy results ranging from 3 days prior to 1 day after the blood culture was drawn.

Performance data were determined for the the first and the second MUG patient cohort using five classifiers with the best performing parametrization trained on the first MUG cohort (Table 3A and B). The least difference in performance was observed using the Neural Network

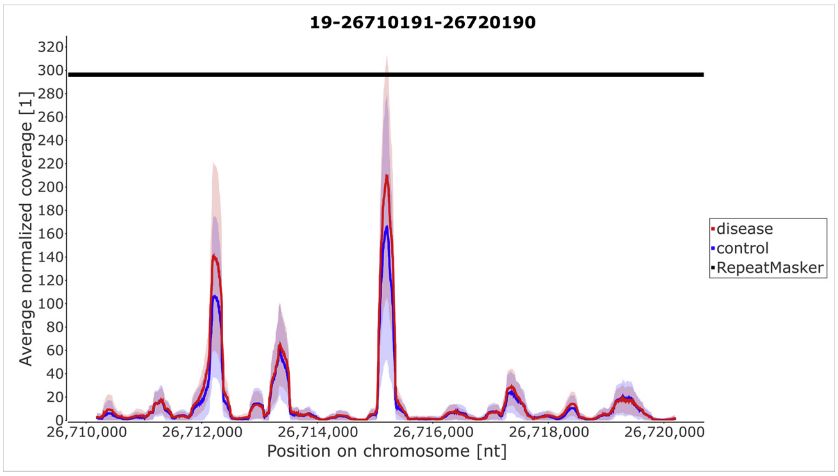


Fig. 2. Mean read coverage and variance of the 10 kilobase pair genomic region surrounding human map locus JC34. The region is located on a centromeric alpha satellite. The mean coverage at the stronger peaks is higher for sepsis patients than for control individuals.

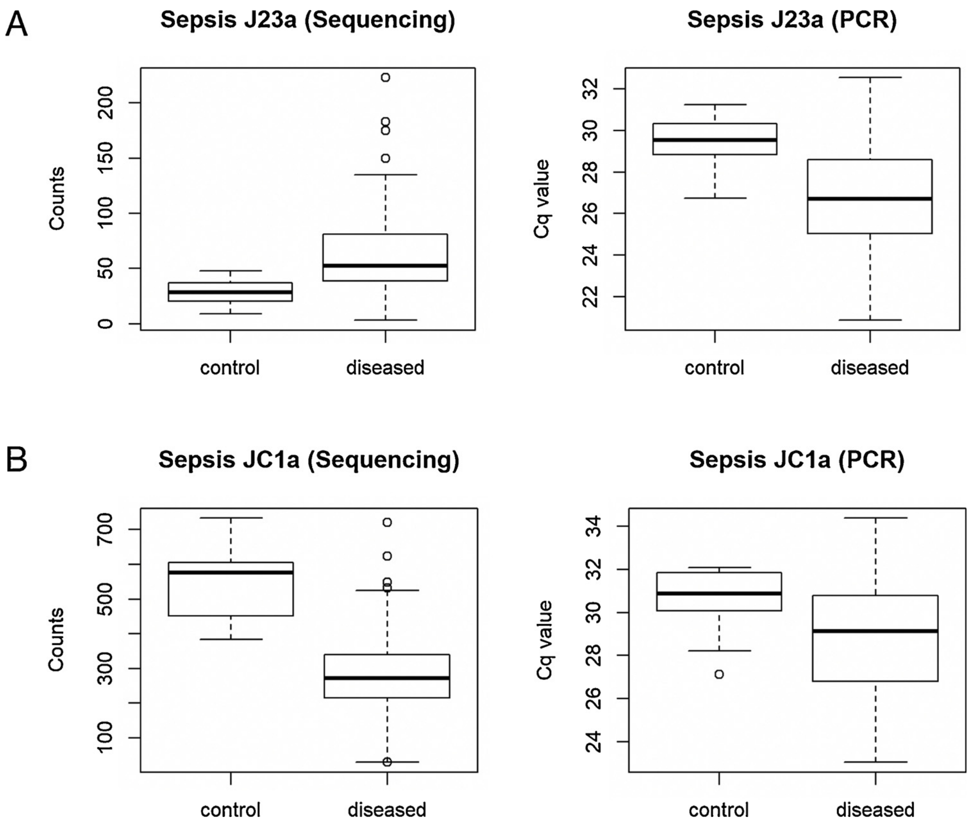


Fig. 3. Sequence mapping results *vs.* RT-PCR results; A. regular relationship (the copy numbers of the motif correlate directly with the number of RT-PCR cycles in patients and controls, respectively.) B. inverse relationship.

(with a difference of 6 % in balanced accuracy) and the Support Vector Machines with Class Weights (with a difference of 20.6 % in balanced accuracy). Concerning the balanced accuracy results across multiple timepoints (Table 4A and B), the best stability across time was observed using Support Vector Machines with Class Weights, showing a difference between maximum and minimum balanced accuracy of 0 % and 3.9 % in First and Second MUG Cohort, respectively.

4. Discussion

Using serum or plasma, as the substrate from which the DNA (CNA) for the high-throughput DNA sequencing was isolated, resulted in genomic read maps, in which regions with peaks in mapping counts, as well as large genomic regions that were essentially not covered by reads, could be distinguished. We conclude that we did not observe any substantial chromosomal DNA contamination, *e.g.* from ruptured nucleated blood cells, as this would have resulted in random, more or less equal coverage along the genomic maps. What we were able to observe were mostly DNA fragments that were protected from exonuclease activity, *i.e.* those fragments that were either packaged into exosomes, or other fragments, which were coiled around the nucleosomes. The fact that we were able to observe distinct and reproducible differences for the copy numbers of some genomic fragments between sepsis patients and controls lets us conclude that indeed the host response (*i.e.* differences in gene network activity) is reflected in the CNA, which is circulating in the human blood stream. This finding is well aligned with other work, especially that of Sadeh et al. (Sadeh et al., 2019).

We were able to identify genomic regions, which did show

Table 3

: MUG Patient Cohort Classification Results (arranged according to the “balanced accuracy” classifier - from highest to lowest). A. First patient cohort, B. Second patient cohort.

Classifier trained	Sensitivity [%]	Specificity [%]	Positive Predictive Value [%]	Negative Predictive Value [%]	Balanced accuracy [%]
A. First MUG Patient Cohort					
Localized Linear Discriminant Analysis	100.0	100.0	100.0	100.0	**100.0**
Support Vector Machines with Class Weights	100.0	100.0	100.0	100.0	**100.0**
Self-Organizing Map	99.3	98.3	99.3	98.3	**98.8**
Mixture Discriminant Analysis	98.5	93.1	97.1	96.4	**95.8**
Neural Network	92.7	81.4	92.1	82.9	**87.1**
B. Second MUG Patient Cohort					
Neural Network	92.9	69.2	89.7	77.1	**81.1**
Support Vector Machines with Class Weights	97.3	61.5	88.0	88.9	**79.4**
Self-Organizing Maps	95.4	48.6	85.1	77.3	**72.0**
Mixture Discriminant Analysis	91.2	48.7	83.7	65.5	**69.9**
Localized Linear Discriminant Analysis	85.8	46.2	82.2	52.9	**66.0**

Table 4
MUG patient cohort "balanced accuracy classifier" results across time points (from 3 days prior to 1 day after the blood culture (*i.e.* day +1) was ordered). Classifiers are arranged as from Table 3B (performance in Second MUG Cohort). A. First patient cohort, B. Second patient cohort.

Classifier trained	Day -3	Day -2	Day -1	Day +1	Day +2
A: First MUG Patient Cohort					
Support Vector Machines with Class Weights	100.0	100.0	100.0	100.0	100.0
Localized Linear Discriminant Analysis	100.0	100.0	100.0	100.0	100.0
Self-Organizing Maps	99.1	99.1	99.1	99.1	98.2
Mixture Discriminant Analysis	96.6	96.6	92.7	95.5	96.6
Neural Network	83.2	84.0	87.6	85.3	88.1
B. Second MUG Patient Cohort					
Neural Network	80.8	81.3	87.4	80.4	81.4
Support Vector Machines with Class Weights	80.8	76.9	80.8	78.0	80.8
Self-Organizing Maps	74.3	74.3	74.3	71.3	70.8
Mixture Discriminant Analysis	62.8	70.5	71.0	70.2	74.4
Localized Linear Discriminant Analysis	57.7	69.2	63.1	64.7	69.9

significant differences in fragment coverage, when comparing sepsis patients and control individuals, thus allowing us to use these particular genomic regions as targets for our qPCR assay development. For most of these marker regions, the results of the high-throughput DNA sequencing experiments and the qPCR experiments showed the same trends, *i.e.* the fragment coverage on the genomic map of healthy and diseased individuals (an thus the copy number of the marker regions in the blood stream) could also be reproduced by the qPCR experiments. This was not true in all instances, especially not for markers mapping within the ribosomal RNA gene regions of the human genome. We suspect that the strong secondary structures of these fragments may have had an influence on the outcome of qPCR experiments, essentially disabling the primer binding or the PCR extension step through the DNA folding.

From our results, it is clear that qPCR with the cell-free fraction of blood as the specimen can be used to detect the onset of human sepsis, allowing to discriminate between patients developing sepsis caused by bacteremia or fungemia and non-sepsis conditions (*e.g.* influenza, pre-eclampsia and lymphoma, as well as healthy controls) with higher accuracy than the currently available molecular sepsis tests. Our classifiers are using pair-wise comparisons of all 23 possible pair-wise combinations of the 24 markers, thus dealing with relative *Cq* values, rather than absolutes. This approach has the advantage that it allows for a wide latitude of DNA concentrations in the cell-free blood fraction, as the ratios are preserved even when a different serum or plasma collection method is used (data of the cattle experiment not shown).

The five best-performing algorithms resulted in a balanced accuracy greater than 87 % each, when used for classification within the sample set used for marker development starting two days before the SEPSIS-3 criteria were met and continuing for at least two additional days. With regards to the second MUG cohort, as is expected for this type of blinded testing on samples, which were not included in the training of the classifier, the balanced accuracy dropped by at least 10 % for these algorithms, thus showing examples of possible drop in performace on validation on an unseen sample set. These overall results outperform the previously published data of any molecular sepsis test that is currently commercially available, for example the Procalcitonin or RNA based assays (Chambliss et al., 2019; Sinha et al., 2018), by more than 10 % in balanced accuracy, even when using our validation results. The detection of the onset of sepsis with two to three days before the first clinical signs is unique to our approach, as none of the other tests has been shown to detect the onset of human sepsis prior to the first clinical signs. This early detection capability needs to be confirmed by employing more samples from days before symptoms (especially Days -3 and -2) in a future study, as the number of available samples was relatively small in the current study. A detection of the presence of human sepsis two days before the SEPSIS-3 criteria are met is a major advantage over the current commercially available tests, as it will allow for an earlier intervention, thus lowering the impact of the disease on the patient and improving the quality of life for sepsis survivors.

Another major advantage of our approach (being based on the host response to the infection), in comparison to tests based on the detection of the infectious organisms, for example *via* the T2 Biosystems system (T2 Biosystems, 2020), which are based on particular DNA sequences of a limited set pathogenic strains, is that our assay can be used to capture all of the bacterial and fungal sepsis cases alike, without having to switch the detection targets. In a recent study, T2Candida revealed discordant Candida species identification in two candidemic patients (Zurl et al., 2019). Six of 22 (27.3 %) deep-seated interstitial cystitis (IC) patients had a positive T2Candida result. Despite advanced time-to-results the clinical value of T2Candida in diagnosing candidemia seemed to be limited by missing blood-culture positive cases. Positivity rates of T2Candida did increase when serial T2Candida samples were tested. The authors also described four cases which occurred due to Candida species outside the T2Candida panel during an invasive candidiasis study period. Due to the predefined panel of detectable Candida species, these cases would have been missed in clinical routine by application of T2Candida testing. Actually Candida auris is not included in the T2Candida panel and might also be missed. A novel T2 identifcation system for Candida auris is now offered, but the product is available for research use only and is not cleared for diagnostic use.

Spencer et al. submitted a patent application in 2015 (Spencer et al., 2015), which describes a method to determine the onset of sepsis using gene expression of a selected set of 266 messenger RNA (mRNA) molecules or selected subsets thereof. Using a subset of 44 of the above mRNA molecules in their assay, this group reached similar performance levels (when compared to the performace of our assay) for the detection of the onset of human sepsis on day -2 (*i.e.* two days before a blood culture was performed), but the performance of the mRNA assay dropped considerably after peaking on day -2. While these results show that it is indeed possible to detect the onset of human sepsis earlier than currently implemented, the performance of our assay remains more stable over the sample period, indicating that the genetic networks used as the basis for the mRNA assay probably changed more rapidly than the host response that our assay is based on. In our opinion, tests based on mRNA would be much more complicated to implement (as they would require an mRNA to DNA conversion step) than our approach, which can directly use cell-free DNA in the PCR assay.

The use of a single DNA-based marker set that can discriminate between patients with both early stages of human sepsis and already presenting clinical signs of sepsis *vs.* patients with other diseases (including those mimicking sepsis) and healthy individuals is to our knowledge unprecedented. In comparison to existing tests or assays, using for instance mRNA, our approach has additional benefits, as it utilizes serum/plasma directly in the qPCR assay. This eliminates the need for steps such as extraction of nucleic acids before testing, or their conversion from RNA to DNA, before a measurement can be performed, thus saving valuable time (when compared to the current standard procedure using blood cultures, which may take one or more additional days) and lowering the cost of a test delivery considerably. With our current marker set, the qPCR assay can already be delivered at a very low cost (we estimate the production cost to be well below 30 US$). The consumables required for our approach are restricted to plasticware and PCR-specific materials (PCR buffer, oligonucleotides and DNA polymerase). The PCR machine used in our experiments was an off-the shelf, commercially available device, without any alterations to either hardware or software, thus making the implementation in hospitals or blood laboratories relatively straightforward, when compared to the establishment of a test that includes a new hardware and software installation. It is important to note that we determine relative *Cq* values (*i.e.* the ratios between markers within a sample) and not absolute *Cq* values (*i.e.* the number of qPCR cylces for each marker in each sample) in our approach. This contributes greatly to the reproducibility of our assay, as issues with varying sampling conditions or storage conditions are minimized due to the preservation of the qPCR *Cq* value ratios in the sampling and testing conditions studied thus far.

5. Conclusion

The presented results are promising for the development of a future commercial assay. We have used a set of 24 markers in this study, thus requiring five wells in a qPCR multiplex assay, when using the BioRad CFX96 Touch™ Real-Time PCR Detection System (or similar qPCR machine types), thus allowing more than ten diagnostic tests to take place simultaneously on a single plate. Given the high sensitivity and specificity values (above 90 %) reached in the first cohort and a modest drop in those values on validation on the second cohort, we are conducting additional studies involving collection of samples from U.S. hospitals to develop a commercial assay for the early detection of sepsis patients. As our approach is based on a generic system in mammalian species, it could also be used to develop diagnostic assays based on host-response markers for many other diseases.

CRediT authorship contribution statement

Elisabeth Ullrich: Conceptualization, Methodology, Writing - original draft. **Petra Heidinger:** Conceptualization, Methodology, Writing - original draft. **Jung Soh:** Software, Writing - review & editing. **Laura Villanova:** Methodology, Software, Validation. **Stefan Grabuschnig:** Software, Validation. **Thorsten Bachler:** Methodology, Validation. **Elisabeth Hirschböck:** Methodology, Validation. **Sara Sánchez-Heredero:** Data curation. **Barry Ford:** Conceptualization, Funding acquisition, Resources. **Maria Sensen:** Methodology, Writing - review & editing. **Ingund Rosales Rodriguez:** Validation, Writing - review & editing. **Daniel Schwendenwein:** Methodology, Validation. **Peter Neumeister:** Resources. **Christoph J. Zurl:** Resources. **Robert Krause:** Conceptualization, Funding acquisition, Resources, Supervision. **Johannes Lorenz Khol:** Resources. **Christoph W. Sensen:** Project administration, Supervision, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgements

This work was partially supported by Defence Research and Development Canada (DRDC) [award number W7702-175822 (Advanced Development of Biomarkers of Sepsis)]. The NOBIS/NOBICS cohort was in part supported by the Austrian Science Fund (FWF) [grant number KLI 561-B31]. This work has been supported by the Federal Ministry for Digital and Economic Affairs (BMDW), the Federal Ministry for Transport, Innovation and Technology (BMVIT), the Styrian Business Promotion Agency (SFG), the Standortagentur Tirol, Government of Lower Austria and ZIT - Technology Agency of the City of Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency (FFG).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jbiotec.2020.01.013.

References

Abu-Saleh, R., Nitzan, O., Saliba, W., Colodner, R., Keness, Y., Yanovskay, A., Edelstein, H., Schwartz, N., Chazan, B., 2018. Bloodstream infections caused by contaminants: epidemiology and risk factors: a 10-Year surveillance. Isr. Med. Assoc. J. 20, 433–437.

Ahmed, A.I., Soliman, R.A., Samir, S., 2016. Cell Free DNA and Procalcitonin as Early Markers of Complications in ICU Patients with Multiple Trauma and Major Surgery. Clin. Lab. 62, 2395–2404. https://doi.org/10.7754/Clin.Lab.2016.160615.

Aucamp, J., Bronkhorst, A.J., Badenhorst, C.P.S., Pretorius, P.J., 2018. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biol. Rev. Camb. Philos. Soc. 93, 1649–1683. https://doi.org/10.1111/brv.12413.

Bender, W.R., Koelper, N.C., Sammel, M.D., Dugoff, L., 2019. Association of fetal fraction of cell-free DNA and hypertensive disorders of pregnancy. Am. J. Perinatol. 36, 311–316. https://doi.org/10.1055/s-0038-1667374.

Blevins, S.M., Bronze, M.S., 2010. Robert Koch and the "golden age" of bacteriology. Int. J. Infect. Dis. 14, e744–51. https://doi.org/10.1016/j.ijid.2009.12.003.

Burnham, P., Khush, K., De Vlaminck, I., 2017. Myriad applications of circulating cell-free DNA in precision organ transplant monitoring. Ann. Am. Thorac. Soc. 14, S237–S241. https://doi.org/10.1513/AnnalsATS.201608-634MG.

Chambliss, A.B., Hayden, J., Colby, J.M., 2019. Evaluation of procalcitonin immunoassay concordance near clinical decision points. Clin. Chem. Lab. Med. https://doi.org/10.1515/cclm-2018-1362.

Cheung, A.H.-K., Chow, C., To, K.-F., 2018. Latest development of liquid biopsy. J. Thorac. Dis. 10, S1645–S1651. https://doi.org/10.21037/jtd.2018.04.68.

Clancy, C.J., Nguyen, M.H., 2013. Finding the "missing 50%" of invasive candidiasis: how nonculture diagnostics will improve understanding of disease spectrum and transform patient care. Clin. Infect. Dis. 56, 1284–1292. https://doi.org/10.1093/cid/cit006.

De Rubis, G., Rajeev Krishnan, S., Bebawy, M., 2019. Liquid Biopsies in Cancer Diagnosis, Monitoring, and Prognosis. Trends Pharmacol. Sci. 40, 172–186. https://doi.org/10.1016/j.tips.2019.01.006.

Duvvuri, B., Lood, C., 2019. Cell-free DNA as a biomarker in autoimmune rheumatic diseases. Front. Immunol. 10, 502. https://doi.org/10.3389/fimmu.2019.00502.

Fenollar, F., Raoult, D., 2007. Molecular Diagnosis of Bloodstream Infections Caused by Non-cultivable Bacteria. https://doi.org/10.1016/j.ijantimicag.2007.06.024.

Fleischmann, C., Scherag, A., Adhikari, N.K.J., Hartog, C.S., Tsaganos, T., Schlattmann, P., Angus, D.C., Reinhart, K., 2016. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am. J. Respir. Crit. Care Med. 193, 259–272. https://doi.org/10.1164/rccm.201504-0781OC.

Gerson, K.D., Truong, S., Haviland, M.J., O'Brien, B.M., Hacker, M.R., Spiel, M.H., 2019. Low fetal fraction of cell-free DNA predicts placental dysfunction and hypertensive disease in pregnancy. Pregnancy Hypertens. 16, 148–153. https://doi.org/10.1016/j.preghy.2019.04.002.

Gielis, E.M., Ledeganck, K.J., Dendooven, A., Meysman, P., Beirnaert, C., Laukens, K., De Schrijver, J., Van Laecke, S., Van Biesen, W., Emonds, M.-P., De Winter, B.Y., Bosmans, J.-L., Del Favero, J., Abramowicz, D., 2019. The use of plasma donor-derived, cell-free DNA to monitor acute rejection after kidney transplantation. Nephrol. Dial. Transplant. https://doi.org/10.1093/ndt/gfz091.

Glebova, K.V., Veiko, N.N., Nikonov, A.A., Porokhovnik, L.N., Kostuyk, S.V., 2018. Cell-free DNA as a biomarker in stroke: current status, problems and perspectives. Crit. Rev. Clin. Lab. Sci. 55, 55–70. https://doi.org/10.1080/10408363.2017.1420032.

Gogenur, M., Burcharth, J., Gogenur, I., 2017. The role of total cell-free DNA in predicting outcomes among trauma patients in the intensive care unit: a systematic review. Crit. Care 21, 14. https://doi.org/10.1186/s13054-016-1578-9.

Grabuschnig, S., Soh, J., Heidinger, P., Bachler, T., Hirschböck, E., Rodriguez, I.R., Schwendenwein, D., Sensen, C.W., 2019. Circulating nucleic acids are predominantly composed of retrotransposable elements and non-telomeric satellite DNA. J. Biotechnol.

Gupta, S., Faap, M., Sakhuja, A., Fasn, M.F., Kumar, G., Mcgrath, E., Nanchal, R.S., Kashani, K.B., Fccp, F., 2016. Culture negative severe Sepsis – nationwide trends and outcomes. Chest. https://doi.org/10.1016/j.chest.2016.08.1460.

Hastie, T., Tibshirani, R., 1996. Discriminant analysis by gaussian mixtures. J. R. Stat. Soc. Ser. B 1, 155–176.

Jackson Chornenki, N.L., Coke, R., Kwong, A.C., Dwivedi, D.J., Xu, M.K., McDonald, E., Marshall, J.C., Fox-Robichaud, A.E., Charbonney, E., Liaw, P.C., 2019. Comparison of the source and prognostic utility of cfDNA in trauma and sepsis. Intensive Care Med. Exp. 7, 29. https://doi.org/10.1186/s40635-019-0251-4.

Jiang, J., Chen, X., Sun, L., Qing, Y., Yang, X., Hu, X., Yang, C., Xu, T., Wang, J., Wang, P., He, L., Dong, C., Wan, C., 2018. Analysis of the concentrations and size distributions of cell-free DNA in schizophrenia using fluorescence correlation spectroscopy. Transl. Psychiatry 8, 104. https://doi.org/10.1038/s41398-018-0153-3.

Karl, P., Erdmann, H.O.M.F., 1896. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. Philos. Trans. R. Soc. London. Ser. A, Contain. Pap. a Math. or Phys. Character 187, 253–318. https://doi.org/10.1098/rsta.1896.0007.

Kohonen, T., Schroeder, M., Huang, T., 2001. Self-Organizing Maps.

Kondo, Y., Umemura, Y., Hayashida, K., Hara, Y., Aihara, M., Yamakawa, K., 2019. Diagnostic value of procalcitonin and presepsin for sepsis in critically ill adult patients: a systematic review and meta-analysis. J. Intensive Care 7, 22. https://doi.org/10.1186/s40560-019-0374-4.

Krause, R., n.d. NOBIS/NOBICS cohort, [WWW Document]. https://clinicaltrials.gov/ct2/show/NCT02801682?term=krause+robert&rank=3 (accessed 1.17.20).

Krause, R., Haberl, R., Wolfler, A., Daxbock, F., Auner, H.W., Krejs, G.J., Wenisch, C., Reisinger, E.C., 2003. Molecular typing of coagulase-negative staphylococcal blood and skin culture isolates to differentiate between bacteremia and contamination. Eur. J. Clin. Microbiol. Infect. Dis. 22, 760–763. https://doi.org/10.1007/s10096-003-1005-4.

Kumar, N., Singh, A.K., 2019. Cell-free fetal DNA: a novel biomarker for early prediction of pre-eclampsia and other obstetric complications. Curr. Hypertens. Rev. 15, 57–63. https://doi.org/10.2174/1573402114666180516131832.

Leon, S.A., Green, A., Yaros, M.J., Shapiro, B., 1975. Radioimmunoassay for nanogram quantities of DNA. J. Immunol. Methods 9, 157–164.

Lindqvist, D., Wolkowitz, O.M., Picard, M., Ohlsson, L., Bersani, F.S., Fernstrom, J., Westrin, A., Hough, C.M., Lin, J., Reus, V.I., Epel, E.S., Mellon, S.H., 2018. Circulating cell-free mitochondrial DNA, but not leukocyte mitochondrial DNA copy number, is elevated in major depressive disorder. Neuropsychopharmacology 43, 1557–1564. https://doi.org/10.1038/s41386-017-0001-9.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17 (No 1) Next Gener. Seq. Data Anal. - 10.14806/ej.17.1.200.

Nguyen, M.H., Clancy, C.J., Pasculle, A.W., Pappas, P.G., Alangaden, G., Pankey, G.A., Schmitt, B.H., Rasool, A., Weinstein, M.P., Widen, R., Hernandez, D.R., Wolk, D.M., Walsh, T.J., Perfect, J.R., Wilson, M.N., Mylonakis, E., 2019. Performance of the T2Bacteria panel for diagnosing bloodstream infections. Ann. Intern. Med. https://doi.org/10.7326/M18-2772.

Oellerich, M., Walson, P.D., Beck, J., Schmitz, J., Kollmar, O., Schutz, E., 2016. Graft-derived cell-free DNA as a marker of transplant graft injury. Ther. Drug Monit. 38 (Suppl 1), S75–S79. https://doi.org/10.1097/FTD.0000000000000239.

Peker, N., Couto, N., Sinha, B., Rossen, J.W., 2018. Diagnosis of bloodstream infections from positive blood cultures and directly from blood samples: recent developments in molecular approaches. Clin. Microbiol. Infect. 24, 944–955. https://doi.org/10.1016/j.cmi.2018.05.007.

Phua, J., Ngerng, W.J., See, K.C., Tay, C.K., Kiong, T., Lim, H.F., Chew, M.Y., Yip, H.S., Tan, A., Khalizah, H.J., Capistrano, R., Lee, K.H., Mukhopadhyay, A., 2013. Characteristics and Outcomes of Culture-negative Versus Culture-positive Severe Sepsis.

Poulet, G., Massias, J., Taly, V., 2019. Liquid biopsy: general concepts. Acta Cytol. 1–7. https://doi.org/10.1159/000499337.

Prost, N., De Razazi, K., Brun-buisson, C., 2013. Unrevealing culture-negative severe sepsis. Crit. Care 1.

Renga, B., 2018. Non invasive prenatal diagnosis of fetal aneuploidy using cell free fetal DNA. Eur. J. Obstet. Gynecol. Reprod. Biol. 225, 5–8. https://doi.org/10.1016/j.ejogrb.2018.03.033.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511812651.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahe, F., 2016. VSEARCH: a versatile open source tool for metagenomics. Peer J. 4, e2584. https://doi.org/10.7717/peerj.2584.

R-project [WWW Document], n.d. https://www.r-project.org/.

Sadeh, R., Fialkoff, G., Sharkia, I., Rahat, A., Nitzan, M., Fox-Fisher, I., Neiman, D., Meler, G., Kamari, Z., Yaish, D., Abu-Gazala, S., Kaplan, T., Shemer, R., Planer, D., Zick, A., Galun, E., Glaser, B., Dor, Y., Friedman, N., 2019. ChIP-seq of plasma cell-free nucleosomes identifies cell-of-origin gene expression programs. bioRxiv 638–643. https://doi.org/10.1101/638643.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., Fulton, R.S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K.M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J.T., Threadgold, G., Torrance, J., Wood, J.M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A.M., Durbin, R., Wilson, R.K., Flicek, P., Eichler, E.E., Church, D.M., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864. https://doi.org/10.1101/gr.213611.116.

Seckel, M., 2017. Sepsis-3: the new definitions. Nurs. Crit. Care 12, 37–43.

Sensen, C.W., Soh, J., Heidinger, P., Villanova, L., Grabuschnig, S., 2019. Methods for treating and detecting Sepsis in humans. Appl. number 16, 526–923.

Sinha, M., Jupe, J., Mack, H., Coleman, T.P., Lawrence, S.M., Fraley, S.I., 2018. Emerging technologies for molecular diagnosis of Sepsis. Clin. Microbiol. Rev. 31. https://doi.org/10.1128/CMR.00089-17.

Spencer, P.M., Lukaszewski, R.A., Craddock, L., Jones, H.E., 2015. Patent. WO2015/121605A1 63. .

Stewart, C.M., Tsui, D.W.Y., 2018. Circulating cell-free DNA for non-invasive cancer management. Cancer Genet. 228–229, 169–179. https://doi.org/10.1016/j.cancergen.2018.02.005.

Stewart, C.M., Kothari, P.D., Mouliere, F., Mair, R., Somnay, S., Benayed, R., Zehir, A., Weigelt, B., Dawson, S.-J., Arcila, M.E., Berger, M.F., Tsui, D.W., 2018. The value of cell-free DNA for molecular pathology. J. Pathol. 244, 616–627. https://doi.org/10.1002/path.5048.

T2 Biosystems [WWW Document], n.d. https://www.t2biosystems.com/homepage-ous/ (accessed 1.17.20).

Thierry, A.R., El Messaoudi, S., Gahan, P.B., Anker, P., Stroun, M., 2016. Origins, structures, and functions of circulating DNA in oncology. Cancer Metastasis Rev. 35, 347–376. https://doi.org/10.1007/s10555-016-9629-x.

Thurairajah, K., Briggs, G.D., Balogh, Z.J., 2018. The source of cell-free mitochondrial DNA in trauma and potential therapeutic strategies. Eur. J. Trauma Emerg. Surg. 44, 325–334. https://doi.org/10.1007/s00068-018-0954-3.

Torio, C.M., Moore, B.J., 2016. Statistical brief #204 national inpatient hospital costs: the most expensive conditions by payer, 2013. Hcup 204, 1–15. https://doi.org/10.1377/hlthaff.2015.1194.3.

Trumpff, C., Marsland, A.L., Basualto-Alarcon, C., Martin, J.L., Carroll, J.E., Sturm, G., Vincent, A.E., Mosharov, E.V., Gu, Z., Kaufman, B.A., Picard, M., 2019. Acute psychological stress increases serum circulating cell-free mitochondrial DNA. Psychoneuroendocrinology 106, 268–276. https://doi.org/10.1016/j.psyneuen.2019.03.026.

Truszewska, A., Foroncewicz, B., Paczek, L., 2017. The role and diagnostic value of cell-free DNA in systemic lupus erythematosus. Clin. Exp. Rheumatol. 35, 330–336.

Tutz, G., Binder, H., 2005. Localized classification. Stat. Comput. 15, 155–166. https://doi.org/10.1007/s11222-005-1305-x.

Vajpeyee, A., Wijatmiko, T., Vajpeyee, M., Taywade, O., 2018. Cell free DNA: A Novel Predictor of Neurological Outcome after Intravenous Thrombolysis and/or Mechanical Thrombectomy in Acute Ischemic Stroke Patients. Neurointervention 13, 13–19. https://doi.org/10.5469/neuroint.2018.13.1.13.

Van den Veyver, I.B., 2016. Recent advances in prenatal genetic screening and testing. F1000Research 5, 2591. https://doi.org/10.12688/f1000research.9215.1.

Vapnik, V., 1998. In: Suykens, J.A.K., Vandewalle, J. (Eds.), The Support Vector Method of Function Estimation BT - Nonlinear Modeling: Advanced Black-Box Techniques. Springer US, Boston, MA, pp. 55–85. https://doi.org/10.1007/978-1-4615-5703-6_3.

Verhoeven, J.G.H.P., Boer, K., Van Schaik, R.H.N., Manintveld, O.C., Huibers, M.M.H., Baan, C.C., Hesselink, D.A., 2018. Liquid biopsies to monitor solid organ transplant function: a review of new biomarkers. Ther. Drug Monit. 40, 515–525. https://doi.org/10.1097/FTD.0000000000000549.

World Health Organization, 2018. Sepsis Fact Sheet WHO Newsroom. [WWW Document] https://www.who.int/news-room/fact-sheets/detail/sepsis (accessed 1.17.20).. .

Zurl, C., Prattes, J., Zollner-schwetz, I., Valentin, T., Rabensteiner, J., Wunsch, S., Hoenigl, M., Krause, R., 2019. T2Candida Magnetic Resonance in Patients With Invasive Candidiasis : Strengths and Limitations. pp. 1–7. https://doi.org/10.1093/mmy/myz101.

Table S1a and S1b: Medical Baseline Characteristics of the 1st MUG Cohort; S1a: Sepsis Group, S1b: Control Group; n.d. not determined, n.a. not applicable, CKD chronic kidney disease

Table S1a: Sepsis Group (n = 63)

	Candida species (n=25)	*Staphylococcus aureus* (n=19)	*Escherichia coli* (n=18)	*Staphylococcus epidermidis* (n=1)
age (years) median (range)	66 (19-90)	67 (52-81)	70.5 (18-96)	50
Sex male n (%)	13 (52%)	12 (63.2%)	6 (28.6%)	1 (100%)
Ventilated n (%)	11 (44%) 1 n.d.	1 (5.3%) 2 n.d.	0	0
Vasopressors n (%)	12 (48%)	1 (5.3%)	0	0
platelets count (G/l) mean (±SD)	221.8 (±148.3) 1 n.d.	199 (±126.8) 1 n.d.	156.4 (±74.8) 1 n.d.	12 (haematological malignancy)
bilirubin (mg/dl) mean (±SD)	2.7 (± 4.7) 9 n.d.	1.3 (±1.3) 13 n.d.	1.8 (±1.6) 7 n.d.	0.3
creatinine (mg/dl) mean (±SD)	1.3 (±0.9) 5 n.d.	3.5 (±3.1) 2 n.d. 3 with CKD	1.7 (±1.3) 1 n.d. 1 with CKD	0.65
Glasgow Coma Score	n.d.	n.d.	n.d.	n.d.
SOFA Score ≥2 n (%)	17 (68%)	11 (57.9%)	12 (66.6%)	0

Table S1b: Control Group (n=47)

	Influenza (n=7)	Lymphoma (n=2) 6 samples from different timepoints	Healthy elective surgery (n=25)	Healthy volunteers (n=13)
age [years] median (range)	63 (40-79)	75 (73-77) mean	49 (19-86)	31 (23-58)
Sex male n (%)	5 (71.4%)	2 (100%)	8 (32%)	6 (46.2%)
Ventilated n (%)	0	0	0	0
Vasopressors n (%)	0	0	0	0
platelets count (G/l) mean (±SD)	193.4 (± 94) 2 n.d.	202 (± 52.9) 1 n.d.	n.d.	n.d.
bilirubin (mg/dl) mean (±SD)	1.5 (n.a.) 6 n.d.	0.5 (± 0.3) 1 n.d.	n.d.	n.d.
creatinine (mg/dl) mean (±SD)	2.5 (±3.3) 2 n.d. 1 with CKD	1.0 (± 0.1) 1 n.d.	n.d.	n.d.
Glasgow Coma Score	n.d.	n.d.	n.d.	15
SOFA Score ≥2 n (%)	1 (14.3%)	0	0	0

Table S2: Motif sequences, RT-PCR primer pairs and genomic regions to which the motifs map with at least 95% identity (human genome assembly GRCh38.p12). For repetitive regions, the number of regions with at least 95% identity is listed.

	Final motif ID	Interim motif ID	Patent ID	Length (bp)	Genomic locus matching amplified sequence with highest BLAST score	Number of genomic regions with at least 95% identity	Forward primer	Reverse primer
1	Sepsis J7	Set1Dis07	hu-sep-CNAD-0100	225	1:176751868-176752092	202	TTGTTGCGATAGTTTACTGAGAATG	TGCTGCTCTAAAGACACATGC
2	Sepsis J15	Set1Dis15	hu-sep-CNAD-0108	211	14:37839103-37839313	64	ACATACGTGTGCATGTGTCTT	AAACAACAGATGCTGGAGAGG
3	Sepsis J16	Set1Dis16	hu-sep-CNAD-0109	191	3:105995079-105995269	192	TCAACATAGTATTGGAAGTTCTGG	GTATCCTGAGACTTTGCTGAAG
4	Sepsis J17	Set1Dis17	hu-sep-CNAD-0110	190	19:50096603-50096792	22	GGTCAGGAGTTCGAGACCA	CAGGCTGGAGTGCAGTG
5	Sepsis J19	Set1Dis19	hu-sep-CNAD-0112	140	11:133643605-133643744	199	GAAACCAACAAGAACAAAGACAC	AATTGTGATGTTAGGGTGTCAA
6	Sepsis J21	Set1Dis21	hu-sep-CNAD-0114	191	18:73823090-73823273	3	TTGTTCTTGTGATAGTTTGCTGAG	GAGACATTACTGACAATAGCAAAGAC
7	Sepsis J23	Set1Dis23	hu-sep-CNAD-0116	200	7:78889332-78889531	150	GAGAGATCCACTGTTAGTCTGATGG	CAATCTAGCAAGGCAGGCCAA
8	Sepsis J36	Set2Dis12	hu-sep-CNAD-0129	210	7:91510533-91510741	179	GGATGCATGGCTGGTTCAA	GTGTGGGTTTGTCATAGATAGCTC
9	Sepsis JU2	Set1Ref02	hu-sep-CNAD-0059	220	6:40091905-40092124	50	TCCAATTCTGTGAAGAAAGTCATTG	AAATACCTAGGAATCCACCTTACAA
10	Sepsis JU4	Set1Ref04	hu-sep-CNAD-0061	241	6:124708424-124708664	179	ACCTTGGGCAGTATGGC	GCATTCTTATACACCAACAACAGA
11	Sepsis JU11	Set1Ref15	hu-sep-CNAD-0068	183	10:91428246-91428428	199	TGGTATCAGTACCATGCTGTTT	GCTACCAATGACTTTCTTCACAG
12	Sepsis JC1	Set1Ctl01	hu-sep-CNAD-0001	166	13:16773515-16773680	23	AAACGTCCGCTTGCAGATAC	GCTGTGAAGATTTCGTTGGAAAC
13	Sepsis JC2	Set1Ctl02	hu-sep-CNAD-0002	122	8:43981494-43981615	16	AGTATGCTGCTGTGTACGTTT	CCTTTGTACTGACAGAGCAGTT
14	Sepsis JC4	Set1Ctl04	hu-sep-CNAD-0004	182	14:16689668-16689849	11	TCAGCTAACAGAGGTGGATCT	CTATGAGTTGAATGGAAATATCCGAAAG

15	Sepsis JC5	Set1Ctl05	hu-sep-CNAD-0005	181	19:25791092-25791272	70	CTTGTGGCCTTCGTTGGAAA	ATTGAACTCAAAGCGGCTGAA
16	Sepsis JC6	Set1Ctl06	hu-sep-CNAD-0006	140	13:16414153-16414293	47	GATAGCTGTGAAGATTTCGTTGG	AGCGTTTCAAACCTCTCTAGG
17	Sepsis JC34	Set2Ctl10	hu-sep-CNAD-0030	204	19:26715090-26715292	22	TGGAAACACTCTGTCTGTAAAGT	CTCCACTTGCAAATTCCACAAA
18	Sepsis JC35	Set2Ctl11	hu-sep-CNAD-0031	151	1:122994551-122994701	2	TCTGCGATGTGTGCGTTC	TCTGTCTAGCAGAATATGAAGAAATCC
19	Sepsis JC42	Set2Ctl18	hu-sep-CNAD-0038	210	1:221211977-221212181	18	GTTCAACCATTGTGGAAGACAG	CCAGTCTATTATTGATGGGCATTT
20	Sepsis JC48	Set2Ctl25	hu-sep-CNAD-0044	131	1:143193394-143193524	19	TCGAATGGACTCGAATGGAATAA	TCGATGATGATCACACTGGATTT
21	Sepsis JC50	Set2Ctl27	hu-sep-CNAD-0046	204	X:33442228-33442431	175	CAACAACTCTTCATGCTATAAACTCTC	GGCCAGAACTTCCAACACTAT
22	Sepsis SC2	HSMC2	hu-sep-CNAD-0050	186	16:36853542-36853727	20	GTGGATATTCGGACCTCTTTGA	GCGCTTGAAATCTCCACTTG
23	Sepsis SC5	HMSC7	hu-sep-CNAD-0053	180	MT:2911-3090	1	CCAACGGAACAAGTTACCCTA	CTGGATTACTCCGGTCTGAAC
24	Sepsis SC7	HSMC14	hu-sep-CNAD-0055	162	MT:9939-10100	1	GTAGATGTGGTTTGACTATTTCTGTATG	GGCTAGGAGGGTGTTGATTATT

Table S3: Performance of the 92 classifiers (arranged according to the “balanced accuracy” - from highest to lowest) on the first patient cohort based on 5 repeats of the 5-fold cross validation and a random search across 5 parameters combinations. Performance measures averaged across different parametrizations and cross-validation runs are reported.

Classifier	Function Name	Sensitivity	Specificity	Positive Predicted Value	Negative Predicted Value	Balanced Accuracy
Self-Organizing Maps	xyf	96,24%	85,08%	93,26%	91,34%	90,66%
Support Vector Machines with Polynomial Kernel	svmPoly	96,00%	84,14%	93,37%	90,04%	90,07%
Localized Linear Discriminant Analysis	loclda	89,33%	90,69%	95,71%	78,51%	90,01%
Support Vector Machines with Class Weights	svmRadialWeights	97,19%	82,41%	92,79%	92,64%	89,80%
Mixture Discriminant Analysis	mda	96,00%	83,45%	93,10%	89,96%	89,72%
Support Vector Machines with Radial Basis Function Kernel	svmRadialCost	95,56%	83,45%	93,07%	88,97%	89,50%
Support Vector Machines with Radial Basis Function Kernel	svmRadialSigma	96,00%	82,76%	92,84%	89,89%	89,38%
Monotone Multi-Layer Perceptron Neural Network	monmlp	93,48%	84,48%	93,34%	84,78%	88,98%
Multi-Layer Perceptron	mlp	94,96%	82,41%	92,63%	87,55%	88,69%
Distance Weighted Discrimination with Radial Basis Function Kernel	dwdRadial	94,67%	82,41%	92,61%	86,91%	88,54%
Support Vector Machines with Radial Basis Function Kernel	svmRadial	95,41%	81,38%	92,26%	88,39%	88,39%
Gaussian Process with Polynomial Kernel	gaussprPoly	98,22%	76,90%	90,82%	94,89%	87,56%
Model Averaged Neural Network	avNNet	94,81%	80,00%	91,69%	86,89%	87,41%
Neural Network	nnet	92.74%	81.38%	92.07%	82.87%	87.06%
AdaBoost Classification Trees	adaboost	94,52%	79,31%	91,40%	86,14%	86,91%
Regularized Random Forest	RRF	94,67%	78,28%	91,03%	86,31%	86,47%
Parallel Random Forest	parRF	94,96%	77,93%	90,92%	86,92%	86,45%
Bagged FDA using gCV Pruning	bagFDAGCV	92,59%	79,31%	91,24%	82,14%	85,95%
C5.0	C5.0	93,19%	78,62%	91,03%	83,21%	85,90%
Bagged AdaBoost	AdaBag	94,37%	76,90%	90,48%	85,44%	85,63%
Stochastic Gradient Boosting	gbm	96,00%	74,48%	89,75%	88,89%	85,24%
High Dimensional Discriminant Analysis	hdda	92,44%	76,55%	90,17%	81,32%	84,50%
Random Forest	rf	94,67%	74,14%	89,50%	85,66%	84,40%
Weighted Subspace Random Forest	wsrf	96,44%	72,07%	88,93%	89,70%	84,26%
Quadratic Discriminant Analysis	qda	96,89%	71,03%	88,62%	90,75%	83,96%
k-Nearest Neighbors	kknn	96,44%	71,38%	88,69%	89,61%	83,91%
Distance Weighted Discrimination with Polynomial Kernel	dwdPoly	92,15%	75,52%	89,75%	80,51%	83,83%
Regularized Discriminant Analysis	rda	94,81%	72,76%	89,01%	85,77%	83,79%
eXtreme Gradient Boosting	xgbLinear	95,56%	71,38%	88,60%	87,34%	83,47%
Boosted Classification Trees	ada	96,89%	70,00%	88,26%	90,63%	83,44%
Multi-Layer Perceptron, with multiple layers	mlpML	93,04%	73,79%	89,20%	81,99%	83,42%
Bagged CART	treebag	93,48%	73,10%	89,00%	82,81%	83,29%
Partial Least Squares	pls	92,44%	74,14%	89,27%	80,83%	83,29%
Random Forest	Rborist	94,37%	71,72%	88,60%	84,55%	83,05%
eXtreme Gradient Boosting	xgbDART	95,26%	70,69%	88,32%	86,50%	82,97%
Tree-Based Ensembles	nodeHarvest	94,22%	71,72%	88,58%	84,21%	82,97%
Linear Discriminant Analysis	lda2	90,81%	74,83%	89,36%	77,78%	82,82%

Heteroscedastic Discriminant Analysis	hda	83,56%	81,72%	91,41%	68,10%	82,64%
Penalized Discriminant Analysis	pda	90,52%	74,48%	89,20%	77,14%	82,50%
Linear Discriminant Analysis with Stepwise Feature Selection	stepLDA	91,85%	72,76%	88,70%	79,32%	82,31%
Linear Discriminant Analysis	lda	91,26%	73,10%	88,76%	78,23%	82,18%
Random Forest	ranger	96,44%	67,59%	87,38%	89,09%	82,02%
Flexible Discriminant Analysis	fda	92,59%	71,38%	88,28%	80,54%	81,99%
Sparse Partial Least Squares	spls	91,56%	72,41%	88,54%	78,65%	81,98%
Naive Bayes	naive_bayes	87,70%	76,21%	89,56%	72,70%	81,96%
Single C5.0 Tree	C5.0Tree	88,00%	74,83%	89,06%	72,82%	81,41%
Generalized Linear Model with Stepwise Feature Selection	glmStepAIC	88,15%	74,48%	88,94%	72,97%	81,32%
Partial Least Squares	kernelpls	91,41%	71,03%	88,02%	78,03%	81,22%
Partial Least Squares	simpls	90,67%	71,72%	88,18%	76,75%	81,20%
Neural Networks with Feature Extraction	pcaNNet	89,48%	72,76%	88,43%	74,82%	81,12%
Multivariate Adaptive Regression Spline	earth	89,78%	72,07%	88,21%	75,18%	80,92%
Penalized Logistic Regression	plr	88,74%	73,10%	88,48%	73,61%	80,92%
Naive Bayes	nb	87,85%	73,79%	88,64%	72,30%	80,82%
Boosted Generalized Linear Model	glmboost	90,52%	71,03%	87,91%	76,30%	80,78%
Sparse Distance Weighted Discrimination	sdwd	89,93%	71,38%	87,97%	75,27%	80,65%
Single C5.0 Ruleset	C5.0Rules	87,11%	74,14%	88,69%	71,19%	80,62%
Rotation Forest	rotationForestCp	94,67%	66,55%	86,82%	84,28%	80,61%
Linear Distance Weighted Discrimination	dwdLinear	89,33%	71,72%	88,03%	74,29%	80,53%
Partial Least Squares	widekernelpls	90,37%	70,34%	87,64%	75,84%	80,36%
Support Vector Machines with Linear Kernel	svmLinear2	89,33%	71,38%	87,90%	74,19%	80,36%
Generalized Additive Model using Splines	gamSpline	89,78%	70,69%	87,70%	74,82%	80,23%
Generalized Partial Least Squares	gpls	89,78%	70,69%	87,70%	74,82%	80,23%
Penalized Multinomial Regression	multinom	89,78%	69,66%	87,32%	74,54%	79,72%
Generalized Linear Model	glm	87,41%	70,34%	87,28%	70,59%	78,88%
Bayesian Generalized Linear Model	bayesglm	90,07%	67,59%	86,61%	74,52%	78,83%
Conditional Inference Random Forest	cforest	93,04%	64,48%	85,91%	79,91%	78,76%
Tree Models from Genetic Algorithms	evtree	85,78%	71,72%	87,59%	68,42%	78,75%
Gaussian Process with Radial Basis Function Kernel	gaussprRadial	98,67%	58,28%	84,63%	94,94%	78,47%
Rotation Forest	rotationForest	94,37%	62,41%	85,39%	82,65%	78,39%
Robust Quadratic Discriminant Analysis	QdaCov	96,00%	59,31%	84,60%	86,43%	77,66%
Regularized Logistic Regression	regLogistic	89,93%	64,83%	85,61%	73,44%	77,38%
CART	rpart1SE	89,48%	65,17%	85,67%	72,69%	77,33%
Shrinkage Discriminant Analysis	sda	90,07%	64,48%	85,51%	73,62%	77,28%
Generalized Linear Model with elasticnet regularization	glmnet	90,81%	60,34%	84,20%	73,84%	75,58%
CART or Ordinal Responses	rpartScore	83,11%	67,93%	85,78%	63,34%	75,52%
Multi-Step Adaptive MCP-Net	msaenet	92,15%	58,62%	83,83%	76,23%	75,38%
CART	rpart2	86,22%	64,48%	84,96%	66,79%	75,35%
Support Vector Machines with Linear Kernel	svmLinear	91,70%	58,97%	83,88%	75,33%	75,33%
Cost-Sensitive CART	rpartCost	86,37%	64,14%	84,86%	66,91%	75,25%
Robust Linear Discriminant Analysis	Linda	85,48%	64,83%	84,98%	65,73%	75,15%
eXtreme Gradient Boosting	xgbTree	89,78%	57,93%	83,24%	70,89%	73,85%
Linear Support Vector Machines with Class Weights	svmLinearWeights	90,22%	56,90%	82,97%	71,43%	73,56%

CART	rpart	85,19%	59,66%	83,09%	63,37%	72,42%
Quadratic Discriminant Analysis with Stepwise Feature Selection	stepQDA	92,59%	50,34%	81,27%	74,49%	71,47%
Patient Rule Induction Method	PRIM	73,93%	64,14%	82,75%	51,38%	69,03%
Stabilized Linear Discriminant Analysis	slda	88,74%	43,45%	78,51%	62,38%	66,09%
Sparse Linear Discriminant Analysis	sparseLDA	98,07%	19,66%	73,97%	81,43%	58,86%
Nearest Shrunken Centroids	pam	96,89%	19,31%	73,65%	72,73%	58,10%
Penalized Discriminant Analysis	pda2	98,52%	14,14%	72,76%	80,39%	56,33%
k-Nearest Neighbors	knn	100,00%	2,76%	70,53%	100,00%	51,38%
Stacked AutoEncoder Deep Neural Network	dnn	100,00%	0,00%	69,95%	NA	50,00%
Non-Informative Model	null	100,00%	0,00%	69,95%	NA	50,00%

Chapter 6

Conclusion

The following itemization enlists the most important findings and results obtained during my PhD. More detailed information regarding these major points are provided within the respective chapters of this thesis.

- A framework for the efficient computational analysis of genomic coverages derived from cfDNA sequencing data was established, which is described in Chapter 2. The procedure for the prediction of candidate marker regions is the costliest functionality in regard of computation time because the calculation of coverage distributions requires consecutive loading of the entire data. The analysis of *e.g.* two sets of 100 samples with 10 million reads each requires grossly a few hours, whereas the creation of genomic alignments for the same set requires two to three days when performed on the same machine. Since this mapping procedure takes by far the longest, although being highly optimized, the calculation and analysis of the corresponding coverage data via the framework can be considered as sufficiently efficient.

- Our experiments and analyses showed that cfDNA in human and animal blood is distributed in a very non-uniform fashion over the genome (see Chapters 4 and 5). More than 50% of the reads sequenced from human blood plasma could be mapped to less than 10 percent of the genome. This highly complex distribution featured strongly covered regions of sizes ranging from about 100 bp up to several kbp, which were interleaved by uncovered or sparsely covered regions.

- The composition of cfDNA was analyzed regarding repeat families and non-repetitive sequence elements. We found that RTEs and different species of satellite DNA were significantly overrepresented (see Chapter 4). This over-representation was observed for human blood plasma, bovine blood serum and the supernatant of cancer cell cultures.

- The cfDNA composition in human blood plasma changes in response to a disease condition, which we showed for human sepsis patients in comparison to healthy individuals (see Chapter 4). We observed a pronounced shift in favor of RTEs, especially for Alu repeats. This finding indicates that information about physiological host conditions reflects in the cfDNA composition, which might be exploited for diagnostic procedures.

- Distinct sequence motifs can be utilized as biomarkers for the diagnosis of sepsis. (see Chapter 5). We demonstrated that the relative abundance of certain motifs, measured via qPCR, can be used for the diagnosis of sepsis caused by bacteria or fungi.

- We have published a review article, which highlights the biological mechanisms of origin behind cfDNA and aligns our findings with the scientific literature (see Chapter 3). The article emphasizes the fact that cfDNA is a product of a multitude of mechanisms, which is sometimes neglected be researchers.

The central question raised by our findings is, how the observed complex distribution of cfDNA in vertebrate blood and the corresponding imbalance between sequence elements in its composition arises. Several pre-analytical factors, such as blood collection procedures, usage of serum or plasma as well as storage and sample preparation are known to affect the outcome of cfDNA analyses [34, 85]. Bronkhorst *et al.* 2016 [80] obtained highly similar results to ours when they analyzed the supernatant of cancer cell cultures. This is especially remarkable, since they used a completely different methodology for sample preparation and sequencing. Therefore, it is very unlikely that our results are attributable to the applied techniques for sample and library preparation or sequencing. Thus, biological mechanisms, *i.e.* sequence specific generation or elimination processes, remain as most plausible explanations.

The clearance of cfDNA in blood occurs via nucleases [86, 87, 88], primarily DNase I [88, 89], but most likely also via liver and kidneys, which at least was shown for foreign DNA injected into mice [86, 90, 91, 92]. Yet, it was also suggested that these mechanisms may not be the major elimination pathways [87, 93]. Blood DNase activity was reported to be up-regulated in response to physical exercises [94], but in contrast found to be significantly reduced in cancer patients [95, 96]. Half-lives of blood cfDNA were determined in different contexts, such as haemodialysis [97], treadmill exercises [98, 99], cancer [100] or pregnancies [86]. These ranged from mere minutes, *e.g.* for treadmill exercises and cffDNA in maternal blood, up to several hours, *e.g.* for ctDNA in the blood of cancer patients, but also showed high inter-individual variability [87, 92]. Since the degradation of cfDNA in the bloodstream depends on DNase activity, it is affected through factors reducing the accessibility of cfDNA to DNases [87]. This may occur *e.g.* via association with macromolecular structures such as virtosomes [101], EVs [102, 103, 104, 105, 106] or binding to plasma proteins [107, 108], which preferentially occurs for methylated DNA [87, 108]. Hence, different sub-populations of cfDNA, associated with EVs, macromolecular structures, serum proteins or being methylated, are expected to degrade with different half-life times. My suggestion to determine the influence of degradation on the composition of cfDNA would be a time-series analysis, where DNA extraction is performed in direct succession to blood draw and afterwards for each following hour. If the over-representation of RTEs and satellite DNA was a result of DNA degradation via DNases, this effect should increase over time. In an unpublished small-scale experiment we have sequenced cfDNA from bovine blood serum. In that experiment DNA extraction was performed 30, 60 and 120 minutes after blood draw. At least in this setup the over-representation of RTEs and satellite DNA was already present at the beginning of the time-series and seemingly did not change significantly afterwards.

Although several mechanisms of origin have already been described thus far (see Chapter 3), beside cell-death, NETosis and enucleation of cells from the haematopoietic linage very little is known about the generation of DNA fragments which are actively released by cells. The excretion of cytosolic DNA from cells via exosomes was shown to be an important physiological process for the maintenance of cellular homeostasis by Takahashi *et al.* 2017 [109]. They also showed that the inhibition of exosome secretion led to intracellular accumulation of cytosolic DNA, which in turn activated the ROS-dependent DNA damage response and in further succession resulted in senescence and cell-death. We hypothesized in our article from Chapter 4 that the activity of the LINE-1 encoded endonuclease-reverse transcriptase [110, 111] and replication stress, occurring *e.g.* at centromeric satellite repeats [112], might be a source of cytosolic DNA fragments. The LINE-1 endonuclease-reverse transcriptase was reported to produce RNA-DNA hybrids via reverse transcription at certain mRNAs in enucleated blood platelets [113], which strongly supports our hypothesis. Finding similar cfDNA compositions and distributions aside from the blood stream in cancer cell culture supernatants, which we did for cell lines of placenta choriocarcinoma and prostate cancer and Bronkhorst *et al.* 2016 [80] did for bone osteosarcoma cell lines, also suggests that a selective release may actually be the major cause behind those observations rather than being the result of elimination processes. A potential approach to assess this idea would be analyzing the effects of inhibiting the LINE-1 endonuclease-reverse transcriptase, *e.g.* via efavirenz [114], and DNA replication, *e.g.* via mimosine [115], on the composition of cfDNA in a cell culture model.

The association of cfDNA with EVs, may play a central role in the preservation of a distinct and actively released subpopulation of cfDNA molecules. These are either partially protected from DNase digestion, due to association with EV surfaces, or fully protected by EV enclosure, where it was consistently reported that only about 20 to 25% of the vesicle associated DNA fragments are contained within the lumen of vesicles [104, 116, 117]. The role of EV associated DNA fragments in the total blood cfDNA concentration was mainly studied in connection to physical exhaustion. Treadmill exercise leads to an increased release of cfDNA into the bloodstream [94, 118], which was mainly attributed to cells of the haematopoietic linage [119], more precisely to neutrophil NETosis [120]. Although a simultaneous elevation of EV levels in blood was reported [102, 103, 104], the exercise related increased release of cfDNA apparently occurs independently from EVs, where only a minute fraction of the total blood cfDNA seems to be associated with EVs, as stated by Helmig *et al*. 2015 [117] and Neuberger *et al.* 2021 [104]. Opposing to this Fernando *et al.* 2017 [121] found that up to 90% of the cfDNA in the blood of healthy donors is associated with exosomes. A major difference in those studies, beside treadmill exercises, was the time-point of DNA purification and library preparation. While this procedure was *e.g.* performed immediately after blood draw in the studies of Helmig *et al*. 2015 [117], in Fernando *et al.* 2017 [121] blood draw and sample preparation occurred at different locations with hours in-between. Thus, it can be coarsely differentiated between two major fractions of blood cfDNA. A short-lived fraction, where levels may substantially rise and decay in response to a stimulus, *e.g.* via physical exertion, exists beside a comparatively smaller but substantially more stable fraction, which is largely associated with EVs. Since for all our samples DNA purification and library preparation occurred hours after blood draw, it can safely be assumed that most of the prepared cfDNA corresponds to the latter fraction. In summary, this cfDNA population was most likely actively released via EVs and featured overrepresented fractions of RTEs and satellite DNA, where the variable sizes, shapes and coverage levels of covered regions apparently did not correspond to the nucleosome-footprint pattern observed for cfDNA stemming from apoptotic cells [122]. Moreover, the RTE fraction of this cfDNA population increased significantly in context to a disease condition comprised by post-surgical sepsis. The extent of this increase and our success in diagnosing sepsis via distinct marker sequences, which not only distinguished samples from sepsis patients from healthy probands but also from other disease conditions such as influenza, let us assume that we measured a systemic host response rather than DNA from dying cells of the afflicted tissue. A specific response to distinct disease conditions, where relevant information is represented by a somewhat substantial fraction of a relatively stable cfDNA population is very advantageous for the design of reliable diagnostic assays. In comparison, cfDNA assays applied in cancer diagnostics [42, 123] aim at the identification of somatic variant specific genomic alterations being present only at a very low number of cfDNA molecules [34]. These comprise only a dwindling small fraction of the total blood cfDNA and therefore analyses suffer from noise in measuring procedures and data analysis [34, 63]. Our conclusions regarding a disease specific host response provides fuel to the discussion about a potential function of cfDNA transferred by EVs as messenger in intercellular communication [124, 125]. These were described to affect gene-regulation of recipient cells [124, 125], produce a DNA-mediated bystander effect, *e.g.* from irradiated cells [126, 127, 128, 129], perform T-cell mediated response activation of dendritic cells [130] and also lead to genomic integration of the shuttled cfDNA molecules [131]. On the contrary, transferred DNA fragments are also a source of cellular stress [109] and DNA damage [132], which also could cause negative effects and stress response. In order to asses if a specific message is contained in a transferred cfDNA population, it would be necessary to investigate if negative effects also occur when cfDNA from healthy donor cells is transferred while ruling out other messenger molecules which are contained in EVs [133].

In my personal opinion, there is still a lot of basic research to be performed on the numerous molecular mechanisms involved with different species of cfDNA. Understanding and exploitation of the diagnostic potential of cfDNA in different body-fluids requires awareness of the enormous complexity comprised by the various routes of generation and elimination. Having the knowledge to specifically target cfDNA fractions where the desired information is contained may lead to a much more versatile and precise usage of cfDNA based diagnostic procedures and even open up the possibility for eventual therapeutic applications.

Bibliography

[1] P. Edman. "A method for the determination of amino acid sequence in peptides". In: *Archives of biochemistry* 22.3 (1949), p. 475. ISSN: 0096-9621.

[2] P. Edman and G. Begg. "A Protein Sequenator". In: *European Journal of Biochemistry* 1.1 (1967), pp. 80–91. ISSN: 14321033. DOI: 10.1111/j.1432-1033.1967.tb00047.x.

[3] Joel B. Hagen. "The origins of bioinformatics". In: *Nature Reviews Genetics* 1.3 (2000), pp. 231–236. ISSN: 14710056. DOI: 10.1038/35042090.

[4] Jeff Gauthier et al. "A brief history of bioinformatics". In: *Briefings in Bioinformatics* 20.6 (2019), pp. 1981–1996. ISSN: 14774054. DOI: 10.1093/bib/bby063.

[5] Margaret Oakley Dayhoff and Robert S. Ledley. "Comprotein: A computer program to aid primary protein structure determination". In: *AFIPS Conference Proceedings - 1962 Fall Joint Computer Conference, AFIPS 1962*. 1962, pp. 262–274. DOI: 10.1145/1461518.1461546.

[6] Linus Pauling et al. "Chemical Paleogenetics. Molecular "Restoration Studies" of Extinct Forms of Life". In: *Acta Chemica Scandinavica* 17 supl. (1963), pp. 9–16. ISSN: 0904-213X. DOI: 10.3891/acta.chem.scand.17s-0009.

[7] Emile Zuckerkandl and Linus Pauling. "Molecules as documents of evolutionary history". In: *Journal of Theoretical Biology* 8.2 (1965), pp. 357–366. ISSN: 10958541. DOI: 10.1016/0022-5193(65)90083-4.

[8] Cédric Notredame. "Recent progress in multiple sequence alignment: A survey". In: *Pharmacogenomics* 3.1 (2002), pp. 131–144. ISSN: 14622416. DOI: 10.1517/14622416.3.1.131.

[9] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 00222836. DOI: 10.1016/0022-2836(70)90057-4.

[10] Walter M. Fitch. "An improved method of testing for evolutionary homology". In: *Journal of Molecular Biology* 16.1 (1966), pp. 9–16. ISSN: 00222836. DOI: 10.1016/S0022-2836(66)80258-9.

[11] Lusheng Wang and Tao Jiang. "On the Complexity of Multiple Sequence Alignment". In: *Journal of Computational Biology* 1.4 (1994), pp. 337–348. ISSN: 15578666. DOI: 10.1089/cmb.1994.1.337.

[12] Da Fei Feng and Russell F. Doolittle. "Progressive sequence alignment as a prerequisite to correct phylogenetic trees". In: *Journal of Molecular Evolution* 25.4 (1987), pp. 351–360. ISSN: 00222844. DOI: 10.1007/BF02603120.

[13] Desmond G. Higgins and Paul M. Sharp. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". In: *Gene* 73.1 (1988), pp. 237–244. ISSN: 03781119. DOI: 10.1016/0378-1119(88)90330-7.

[14] Cédric Notredame. “Recent evolutions of multiple sequence alignment algorithms”. In: *PLoS Computational Biology* 3.8 (2007), pp. 1405–1408. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.0030123.

[15] Marshall Nirenberg. “Historical review: Deciphering the genetic code - A personal account”. In: *Trends in Biochemical Sciences* 29.1 (2004), pp. 46–54. ISSN: 09680004. DOI: 10.1016/j.tibs.2003.11.009.

[16] Heinrich Matthaei and Marshall W. Nirenberg. “The dependence of cell-free protein synthesis in E.coli upon RNA prepared from ribosomes”. In: *Biochemical and Biophysical Research Communications* 4.6 (1961), pp. 404–408. ISSN: 10902104. DOI: 10.1016/0006-291X(61)90298-4.

[17] F. H.C. Crick. “The origin of the genetic code”. In: *Journal of Molecular Biology* 38.3 (1968), pp. 367–379. ISSN: 00222836. DOI: 10.1016/0022-2836(68)90392-6.

[18] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467. ISSN: 00278424. DOI: 10.1073/pnas.74.12.5463.

[19] F. Sanger and A. R. Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. In: *Journal of Molecular Biology* 94.3 (1975). ISSN: 00222836. DOI: 10.1016/0022-2836(75)90213-2.

[20] A. M. Maxam and W. Gilbert. “A new method for sequencing DNA”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.2 (1977), pp. 560–564. ISSN: 00278424. DOI: 10.1073/pnas.74.2.560.

[21] W. J.S. Diniz and F. Canduri. “Bioinformatics: An overview and its applications”. In: *Genetics and Molecular Research* 16.1 (2017). ISSN: 16765680. DOI: 10.4238/gmr16019645.

[22] Ziheng Yang and Bruce Rannala. “Molecular phylogenetics: Principles and practice”. In: *Nature Reviews Genetics* 13.5 (2012), pp. 303–314. ISSN: 14710056. DOI: 10.1038/nrg3186.

[23] Mark J.P. Chaisson, Richard K. Wilson, and Evan E. Eichler. “Genetic variation and the de novo assembly of human genomes”. In: *Nature Reviews Genetics* 16.11 (2015), pp. 627–640. ISSN: 14710064. DOI: 10.1038/nrg3933.

[24] Jianmin Wu et al. “KOBAS server: A web-based platform for automated annotation and pathway identification”. In: *Nucleic Acids Research* 34.WEB. SERV. ISS. (2006). ISSN: 03051048. DOI: 10.1093/nar/gkl167.

[25] Tobias Marschall and Sven Rahmann. “Efficient exact motif discovery”. In: *Bioinformatics* 25.12 (2009). ISSN: 13674803. DOI: 10.1093/bioinformatics/btp188.

[26] Ali Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7 (2008), pp. 621–628. ISSN: 15487091. DOI: 10.1038/nmeth.1226.

[27] Stephen Hilgartner. “Biomolecular Databases: New Communication Regimes for Biology?” In: *Science Communication* 17.2 (1995), pp. 240–263. ISSN: 15528545. DOI: 10.1177/1075547095017002009.

[28] Eric W. Sayers et al. “GenBank”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D92–D96. ISSN: 13624962. DOI: 10.1093/nar/gkaa1023.

[29] Peter W. Harrison et al. "The European Nucleotide Archive in 2020". In: *Nucleic Acids Research* 49.D1 (2021), pp. D82–D85. ISSN: 13624962. DOI: 10.1093/nar/gkaa1028.

[30] Helen M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. ISSN: 03051048. DOI: 10.1093/nar/28.1.235.

[31] Stephen F. Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* (1990). ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2.

[32] Chiang Ching Huang, Meijun Du, and Liang Wang. "Bioinformatics analysis for circulating cell-free DNA in cancer". In: *Cancers* 11.6 (2019). ISSN: 20726694. DOI: 10.3390/cancers11060805.

[33] Landon L. Chan and Peiyong Jiang. *Bioinformatics analysis of circulating cell-free DNA sequencing data*. 2015. DOI: 10.1016/j.clinbiochem.2015.04.022.

[34] Zuzana Pös et al. "Technical and methodological aspects of cell-free nucleic acids analyzes". In: *International Journal of Molecular Sciences* (2020). ISSN: 14220067. DOI: 10.3390/ijms21228634.

[35] Lisa Hui, Jill L. Maron, and Peter B. Gahan. "Other body fluids as non-invasive sources of cell-free DNA/RNA". In: *Advances in Predictive, Preventive and Personalised Medicine* 5 (2015), pp. 295–323. ISSN: 22113509. DOI: 10.1007/978-94-017-9168-7_11.

[36] Janine Aucamp et al. "The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature". In: *Biological Reviews* 93.3 (2018), pp. 1649–1683. ISSN: 1469185X. DOI: 10.1111/brv.12413.

[37] P. Mandel and P. Metais. "Les acides nucleiques du plasma sanguin chez l' homme [The nucleic acids in blood plasma in humans]." In: *C R Seances Soc Biol Fil.* 142.3-4 (1948), pp. 241–243.

[38] J. D. Watson and F. H.C. Crick. "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738. ISSN: 00280836. DOI: 10.1038/171737a0.

[39] S. A. Leon et al. "Free DNA in the Serum of Cancer Patients and the Effect of Therapy". In: *Cancer Research* 37.3 (1977), pp. 646–650. ISSN: 15387445.

[40] S. A. Leon et al. "Free DNA in the serum of rheumatoid arthritis patients." In: *Journal of Rheumatology* 4.2 (1977), pp. 139–143. ISSN: 0315162X.

[41] E. M. Tan et al. "Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus." In: *The Journal of clinical investigation* 45.11 (1966), pp. 1732–1740. ISSN: 00219738. DOI: 10.1172/JCI105479.

[42] David Chu and Ben Ho Park. "Liquid biopsy: unlocking the potentials of cell-free DNA". In: *Virchows Archiv* 471.2 (2017), pp. 147–154. ISSN: 14322307. DOI: 10.1007/s00428-017-2137-8.

[43] Jonathan C.M. Wan et al. *Liquid biopsies come of age: Towards implementation of circulating tumour DNA*. 2017. DOI: 10.1038/nrc.2017.7.

[44] Alicia Oshlack, Mark D. Robinson, and Matthew D. Young. "From RNA-seq reads to differential expression results". In: *Genome Biology* (2010). ISSN: 14747596. DOI: 10.1186/gb-2010-11-12-220.

[45] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. "RNA-Seq differential expression analysis: An extended review and a software tool". In: *PLoS ONE* (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0190152.

[46] Francesca Galardi et al. "Cell-free dna-methylation-based methods and applications in oncology". In: *Biomolecules* 10.12 (2020), pp. 1–23. ISSN: 2218273X. DOI: 10.3390/biom10121677.

[47] Matthias Lienhard et al. "MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments". In: *Bioinformatics* (2014). ISSN: 13674803. DOI: 10.1093/bioinformatics/btt650.

[48] Bjørn Fjukstad and Lars Ailo Bongo. "A Review of Scalable Bioinformatics Pipelines". In: *Data Science and Engineering* (2017). ISSN: 23641541. DOI: 10.1007/s41019-017-0047-z.

[49] Jeremy Leipzig. "A review of bioinformatic pipeline frameworks". In: *Briefings in Bioinformatics* (2017). ISSN: 14774054. DOI: 10.1093/bib/bbw020.

[50] David R. Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* (2008). ISSN: 00280836. DOI: 10.1038/nature07517.

[51] Sowmiya Moorthie, Christopher J. Mattocks, and Caroline F. Wright. "Review of massively parallel DNA sequencing technologies". In: (2011). ISSN: 18776558. DOI: 10.1007/s11568-011-9156-3.

[52] Nicholas Caruccio. "Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition". In: *High-Throughput Next Generation Sequencing*. Springer, 2011, pp. 241–255.

[53] G. F. Hong. "A method for sequencing single-stranded cloned DNA in both directions". In: *Bioscience Reports* 1.3 (1981), pp. 243–252. ISSN: 01448463. DOI: 10.1007/BF01114911.

[54] Melissa J. Fullwood et al. "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses". In: *Genome Research* 19.4 (2009), pp. 521–532. ISSN: 10889051. DOI: 10.1101/gr.074906.107.

[55] Peter J.A. Cock et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic Acids Research* (2009). ISSN: 03051048. DOI: 10.1093/nar/gkp1137.

[56] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (2011), p. 10. DOI: 10.14806/ej.17.1.200.

[57] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: A flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu170.

[58] Cristian Del Fabbro et al. "An extensive evaluation of read trimming effects on illumina NGS data analysis". In: *PLoS ONE* (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0085024.

[59] Brent Ewing et al. "Base-calling of automated sequencer traces using phred. I. Accuracy assessment". In: *Genome Research* 8.3 (1998), pp. 175–185. ISSN: 10889051. DOI: 10.1101/gr.8.3.175.

[60] Brent Ewing and Phil Green. “Base-calling of automated sequencer traces using phred. II. Error probabilities”. In: *Genome Research* 8.3 (1998), pp. 186–194. ISSN: 10889051. DOI: 10.1101/gr.8.3.186.

[61] Illumina. “Nextera ® XT DNA Library Preparation Kit”. In: *Reporter* (2014).

[62] M. Mielczarek and J. Szyda. *Review of alignment and SNP calling algorithms for next-generation sequencing data*. 2016. DOI: 10.1007/s13353-015-0292-7.

[63] Philip Burnham et al. “Separating the signal from the noise in metagenomic cell-free DNA sequencing”. In: *Microbiome* 8.1 (2020). ISSN: 20492618. DOI: 10.1186/s40168-020-0793-4.

[64] Heng Li and Nils Homer. *A survey of sequence alignment algorithms for next-generation sequencing*. 2010. DOI: 10.1093/bib/bbq015.

[65] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3 (2009). ISSN: 14747596. DOI: 10.1186/gb-2009-10-3-r25.

[66] Heng Li et al. “The Sequence Alignment / Map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 13674803. DOI: {10.1093/bioinformatics/btp352}.

[67] The SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification*. https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf. 2021.

[68] Ben Langmead and Steven L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359. ISSN: 15487091. DOI: 10.1038/nmeth.1923.

[69] Paolo Ferragina and Giovanni Manzini. “Opportunistic data structures with applications”. In: *Annual Symposium on Foundations of Computer Science - Proceedings*. 2000, pp. 390–398. DOI: 10.1109/sfcs.2000.892127.

[70] Tsunglin Liu et al. “Joining Illumina paired-end reads for classifying phylogenetic marker sequences”. In: *BMC Bioinformatics* (2020). ISSN: 14712105. DOI: 10.1186/s12859-020-3445-6.

[71] Marie Agnès Dillies et al. “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In: *Briefings in Bioinformatics* 14.6 (2013), pp. 671–683. ISSN: 14675463. DOI: 10.1093/bib/bbs046.

[72] Simon Andrews et al. *FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics*. 2015.

[73] Philip Ewels et al. “MultiQC: Summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* (2016). ISSN: 14602059. DOI: 10.1093/bioinformatics/btw354.

[74] Broad Institute. *Picard:A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*. 2016.

[75] Christopher M. Bishop. *Pattern Recoginiton and Machine Learning*. 2006, p. 738. ISBN: 978-0-387-31073-2.

[76] S. Kullback and R. A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951). ISSN: 0003-4851. DOI: 10.1214/aoms/1177729694.

[77] Ronald E. Shiffler. “Maximum z scores and outliers”. In: *American Statistician* (1988). ISSN: 15372731. DOI: 10.1080/00031305.1988.10475530.

[78] Peter J. Rousseeuw and Mia Hubert. “Robust statistics for outlier detection”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2011). ISSN: 19424787. DOI: 10.1002/widm.2.

[79] Jonathan Casper et al. “The UCSC Genome Browser database: 2018 update”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D762–D769. ISSN: 13624962. DOI: 10.1093/nar/gkx1020.

[80] Abel Jacobus Bronkhorst et al. “Characterization of the cell-free DNA released by cultured cancer cells”. In: *Biochimica et Biophysica Acta - Molecular Cell Research* 1863.1 (2016), pp. 157–165. ISSN: 18792596. DOI: 10.1016/j.bbamcr.2015.10.022.

[81] HA Johne and L Frothingham. “Ein eigentuemlicher Fall von Tuberculose beim Rind”. In: *Deutsche Zeitschrift f. Tiermedizin* 21 (1895), pp. 438–454.

[82] Ofelia Chacon, Luiz E. Bermudez, and Raúl G. Barletta. “Johne’s disease, inflammatory bowel disease, and Mycobacterium paratuberculosis”. In: *Annual Review of Microbiology* (2004). ISSN: 00664227. DOI: 10.1146/annurev.micro.58.030603.123726.

[83] Walter Baumgartner and Johannes Khol. “Paratuberculosis (Johne’s disease) in ruminants-an ongoing story”. In: *Slovenian Veterinary Research* 43 (Jan. 2006). ISSN: 15804003.

[84] Philip Rasmussen et al. “Economic losses due to Johne’s disease (paratuberculosis) in dairy cattle”. In: *Journal of Dairy Science* (Jan. 2021). DOI: 10.3168/jds.2020-19381.

[85] Romain Meddeb, Ekaterina Pisareva, and Alain R. Thierry. “Guidelines for the preanalytical conditions for analyzing circulating cell-free DNA”. In: *Clinical Chemistry* (2019). ISSN: 15308561. DOI: 10.1373/clinchem.2018.298323.

[86] Y. M. Dennis Lo et al. “Rapid clearance of fetal DNA from maternal plasma”. In: *American Journal of Human Genetics* (1999). ISSN: 00029297. DOI: 10.1086/302205.

[87] Sonia Khier and Laura Lohan. “Kinetics of circulating cell-free DNA for biomedical applications: Critical appraisal of the literature”. In: *Future Science OA* 4.4 (2018). ISSN: 20565623. DOI: 10.4155/fsoa-2017-0140.

[88] Gustavo Barcelos Barra et al. “EDTA-mediated inhibition of DNases protects circulating cell-free DNA from ex vivo degradation in blood samples”. In: *Clinical Biochemistry* (2015). ISSN: 18732933. DOI: 10.1016/j.clinbiochem.2015.02.014.

[89] Roger M. Herriott, John H. Connolly, and Shanta Gupta. “Blood nucleases and infectious viral nucleic acids”. In: *Nature* (1961). ISSN: 00280836. DOI: 10.1038/189817a0.

[90] T M Chused, A D Steinberg, and N Talal. “The clearance and localization of nucleic acids by New Zealand and normal mice.” In: *Clinical and experimental immunology* (1972). ISSN: 0009-9104.

[91] Woodruff Emlen and Mart Mannik. “Kinetics and mechanisms for removal of circulating single-stranded dna in mice*”. In: *Journal of Experimental Medicine* (1978). ISSN: 15409538. DOI: 10.1084/jem.147.3.684.

[92] Anatoli Kustanovich et al. “Life and death of circulating cell-free DNA”. In: *Cancer Biology and Therapy* 20.8 (2019), pp. 1057–1067. ISSN: 15558576. DOI: 10.1080/15384047.2019.1598759.

[93] Stephanie C.Y. Yu et al. “High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing”. In: *Clinical Chemistry* (2013). ISSN: 00099147. DOI: 10.1373/clinchem.2013.203679.

[94] Martina Velders et al. “Exercise is a potent stimulus for enhancing circulating DNase activity”. In: *Clinical Biochemistry* (2014). ISSN: 18732933. DOI: 10.1016/j.clinbiochem.2013.12.017.

[95] Svetlana N. Tamkovich et al. “Circulating DNA and DNase Activity in Human Blood”. In: *Annals of the New York Academy of Sciences* 1075.1 (2006), pp. 191–196. DOI: https://doi.org/10.1196/annals.1368.026.

[96] Béatrice Dewez et al. “Serum Alkaline Deoxyribonuclease Activity, a Sensitive Marker for the Therapeutic Monitoring of Cancer Patients: Methodological Aspects”. In: *Clinical Chemistry and Laboratory Medicine* (1993). ISSN: 14374331. DOI: 10.1515/cclm.1993.31.11.793.

[97] P. Rumore et al. “Haemodialysis as a model for studying endogenous plasma DNA: Oligonucleosome-like structure and clearance”. In: *Clinical and Experimental Immunology* (1992). ISSN: 00099104. DOI: 10.1111/j.1365-2249.1992.tb05831.x.

[98] Thomas Beiter et al. “Short-term treadmill running as a model for studying cell-free DNA kinetics in vivo”. In: *Clinical Chemistry* (2011). ISSN: 00099147. DOI: 10.1373/clinchem.2010.158030.

[99] Sarah Breitbach et al. “Direct measurement of cell-free DNA from serially collected capillary plasma during incremental exercise”. In: *Journal of Applied Physiology* (2014). ISSN: 15221601. DOI: 10.1152/japplphysiol.00002.2014.

[100] Frank Diehl et al. “Circulating mutant DNA to assess tumor dynamics”. In: *Nature Medicine* (2008). ISSN: 10788956. DOI: 10.1038/nm.1789.

[101] Peter B. Gahan and Maurice Stroun. “The virtosome-a novel cytosolic informative entity and intercellular messenger”. In: *Cell Biochemistry and Function* 28.7 (2010), pp. 529–538. ISSN: 02636484. DOI: 10.1002/cbf.1690.

[102] Alexandra Brahmer et al. “Platelets, endothelial cells and leukocytes contribute to the exercise-triggered release of extracellular vesicles into the circulation”. In: *Journal of Extracellular Vesicles* 8.1 (2019). ISSN: 20013078. DOI: 10.1080/20013078.2019.1615820.

[103] Carsten Frühbeis et al. “Physical exercise induces rapid release of small extracellular vesicles into the circulation”. In: *Journal of Extracellular Vesicles* (2015). ISSN: 20013078. DOI: 10.3402/jev.v4.28239.

[104] Elmo W. I. Neuberger et al. “Kinetics and topology of DNA associated with circulating extracellular vesicles released during exercise”. In: *bioRxiv* (2021). DOI: 10.1101/2021.02.12.430930.

[105] D. W. Dorward, C. F. Garon, and R. C. Judd. “Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae”. In: *Journal of Bacteriology* 171.5 (1989), pp. 2499–2505. ISSN: 00219193. DOI: 10.1128/jb.171.5.2499-2505.1989.

[106] C. F. Garon, D. W. Dorward, and M. D. Corwin. “Structural features of borrelia burgdorferi - The lyme disease spirochete: Silver staining for nucleic acids”. In: *Scanning Microscopy*. Vol. 3. SUPPL. 3. 1989, pp. 109–115.

[107] Ulrich Krach-Hansen, Victor Tuan Giam Chuang, and Masaki Otagiri. “Practical aspects of the ligand-binding and enzymatic properties of human serum albumin”. In: *Biological and Pharmaceutical Bulletin* (2002). ISSN: 09186158. DOI: 10.1248/bpb.25.695.

[108] Tatyana E Skvortsova et al. “Methylated cell-free DNA in vitro and in vivo”. In: *Circulating nucleic acids in plasma and serum*. Springer, 2010, pp. 185–194.

[109] Akiko Takahashi et al. “Exosomes maintain cellular homeostasis by excreting harmful DNA from cells”. In: *Nature Communications* 8 (2017). ISSN: 20411723. DOI: 10.1038/ncomms15287.

[110] Qinghua Feng et al. “Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition”. In: *Cell* 87.5 (1996), pp. 905–916. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)81997-2.

[111] Stephen L. Mathias et al. “Reverse transcriptase encoded by a human transposable element”. In: *Science* 254.5039 (1991), pp. 1808–1810. ISSN: 00368075. DOI: 10.1126/science.1722352.

[112] Zhengke Li et al. “hDNA2 nuclease/helicase promotes centromeric DNA replication and genome stability”. In: *The EMBO Journal* 37.14 (2018), e96729. ISSN: 0261-4189. DOI: 10.15252/embj.201796729.

[113] Hansjörg Schwertz et al. “Endogenous LINE-1 (Long Interspersed Nuclear Element-1) Reverse Transcriptase Activity in Platelets Controls Translational Events Through RNA-DNA Hybrids”. In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 38.4 (2018), pp. 801–815. ISSN: 15244636. DOI: 10.1161/ATVBAHA.117.310552.

[114] Radhika Patnala et al. “Inhibition of LINE-1 retrotransposon-encoded reverse transcriptase modulates the expression of cell differentiation genes in breast cancer cells”. In: *Breast Cancer Research and Treatment* (2014). ISSN: 01676806. DOI: 10.1007/s10549-013-2812-7.

[115] D. M. Gilbert et al. “Mimosine arrests DNA synthesis at replication forks by inhibiting deoxyribonucleotide metabolism”. In: *Journal of Biological Chemistry* (1995). ISSN: 00219258. DOI: 10.1074/jbc.270.16.9597.

[116] Elisa Lázaro-Ibáñez et al. “DNA analysis of low- and high-density fractions defines heterogeneous subpopulations of small extracellular vesicles based on their DNA cargo and topology”. In: *Journal of Extracellular Vesicles* 8.1 (2019). ISSN: 20013078. DOI: 10.1080/20013078.2019.1656993.

[117] Susanne Helmig et al. “Release of bulk cell free DNA during physical exercise occurs independent of extracellular vesicles”. In: *European Journal of Applied Physiology* 115.11 (2015), pp. 2271–2280. ISSN: 14396319. DOI: 10.1007/s00421-015-3207-8.

[118] Johanna Atamaniuk et al. “Increased concentrations of cell-free plasma DNA after exhaustive exercise”. In: *Clinical Chemistry* (2004). ISSN: 00099147. DOI: 10.1373/clinchem.2004.034553.

[119] Suzan Tug et al. “Exercise-induced increases in cell free DNA in human plasma originate predominantly from cells of the haematopoietic lineage”. In: *Exercise Immunology Review* 21 (2015), pp. 164–173. ISSN: 10775552.

[120] Thomas Beiter et al. "Neutrophils release extracellular DNA traps in response to exercise". In: *Journal of Applied Physiology* (2014). ISSN: 15221601. DOI: 10.1152/japplphysiol.00173.2014.

[121] M. Rohan Fernando et al. "New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes". In: *PLoS ONE* 12.8 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0183915.

[122] Matthew W. Snyder et al. "Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin". In: *Cell* 164.1-2 (2016), pp. 57–68. ISSN: 10974172. DOI: 10.1016/j.cell.2015.11.050.

[123] K. C.Allen Chan et al. "Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing". In: *Clinical Chemistry* 59.1 (2013), pp. 211–224. ISSN: 00099147. DOI: 10.1373/clinchem.2012.196014.

[124] Jin Cai et al. "Extracellular vesicle-mediated transfer of donor genomic DNA to recipient cells is a novel mechanism for genetic influence between cells". In: *Journal of Molecular Cell Biology* 5.4 (2013), pp. 227–238. ISSN: 16742788. DOI: 10.1093/jmcb/mjt011.

[125] Jin Cai et al. "Functional transferred DNA within extracellular vesicles". In: *Experimental Cell Research* 349.1 (2016), pp. 179–183. ISSN: 10902422. DOI: 10.1016/j.yexcr.2016.10.012.

[126] Aleksei V. Ermakov et al. "Oxidative stress as a significant factor for development of an adaptive response in irradiated and nonirradiated human lymphocytes after inducing the bystander effect by low-dose X-radiation". In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 669.1-2 (2009), pp. 155–161. ISSN: 00275107. DOI: 10.1016/j.mrfmmm.2009.06.005.

[127] Aleksei V. Ermakov et al. "An extracellular DNA mediated bystander effect produced from low dose irradiated endothelial cells". In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 712.1-2 (2011), pp. 1–10. ISSN: 00275107. DOI: 10.1016/j.mrfmmm.2011.03.002.

[128] Marina S. Konkova et al. "Oxidized Cell-Free DNA Is a Factor of Stress Signaling in Radiation-Induced Bystander Effects in Different Types of Human Cells". In: *International Journal of Genomics* 2019 (2019). ISSN: 23144378. DOI: 10.1155/2019/9467029.

[129] Kentaro Ariyoshi et al. "Radiation-Induced Bystander Effect is Mediated by Mitochondrial DNA in Exosome-Like Vesicles". In: *Scientific Reports* 9.1 (2019). ISSN: 20452322. DOI: 10.1038/s41598-019-45669-z.

[130] Daniel Torralba et al. "Priming of dendritic cells by DNA-containing extracellular vesicles from activated T cells through antigen-driven contacts". In: *Nature Communications* (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-05077-9.

[131] Stefanie Fischer et al. "Indication of horizontal DNA gene transfer by extracellular vesicles". In: *PLoS ONE* 11.9 (2016). ISSN: 19326203. DOI: 10.1371/journal.pone.0163665.

[132] Ranjan Basak, Naveen Kumar Nair, and Indraneel Mittra. “Evidence for cell-free nucleic acids as continuously arising endogenous DNA mutagens”. In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* (2016). ISSN: 18792871. DOI: 10.1016/j.mrfmmm.2016.10.002.

[133] Graça Raposo and Willem Stoorvogel. “Extracellular vesicles: Exosomes, microvesicles, and friends”. In: *Journal of Cell Biology* 200.4 (2013), pp. 373–383. ISSN: 00219525. DOI: 10.1083/jcb.201211138.